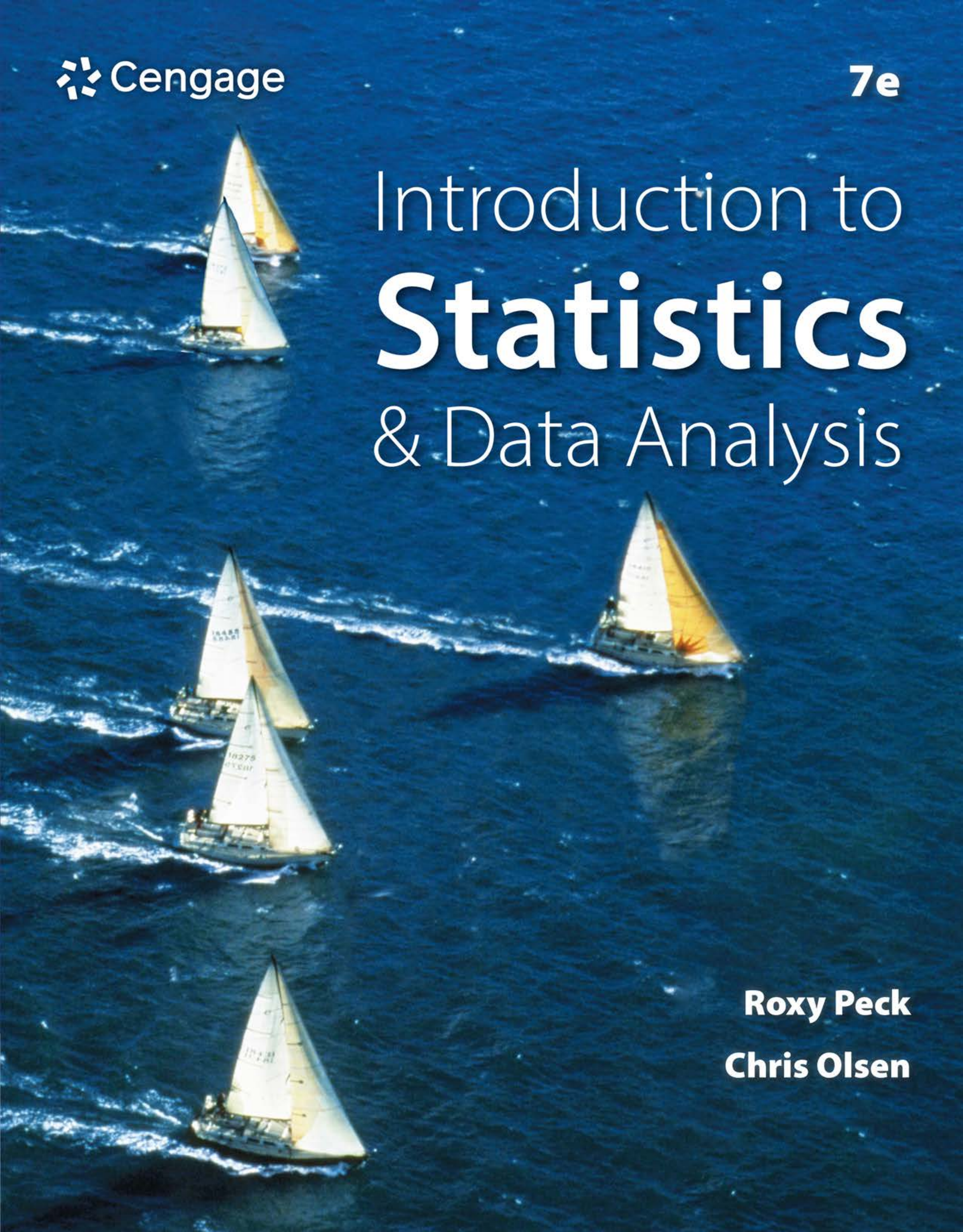


Cengage

7e

Introduction to Statistics & Data Analysis

Roxy Peck

Chris Olsen

Introduction to Statistics and Data Analysis

7th Edition

Roxy Peck
California Polytechnic State University, San Luis Obispo

Chris Olsen
Grinnell College

Cengage

Australia • Brazil • Canada • Mexico • Singapore • United Kingdom • United States

This is an electronic version of the print textbook. Due to electronic rights restrictions, some third party content may be suppressed. Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. The publisher reserves the right to remove content from this title at any time if subsequent rights restrictions require it. For valuable information on pricing, previous editions, changes to current editions, and alternate formats, please visit www.cengage.com/highered to search by ISBN#, author, title, or keyword for materials in your areas of interest.

Important Notice: Media content referenced within the product description or the product text may not be available in the eBook version.

Cengage

Introduction to Statistics and Data Analysis,
Seventh Edition
Roxy Peck and Chris Olsen

Portfolio Product Director: Rita Lombard

Portfolio Product Manager: Adrian Daniel

Sr. Learning Designer: Elinor Gregory

Sr. Product Assistant: Amelia Padilla

Lead Marketing Manager: Tom Ziolkowski

Content Manager: Lynh Pham

Content Acquisition Analyst: Andrea White

Production Service: MPS Limited

Compositor: MPS Limited

Text and Cover Designer: Tim Biddick

Cover Image: Charles O'Rear/Corbis Documentary/Getty images

Copyright © 2025 Cengage Learning, Inc. ALL RIGHTS RESERVED.
WCN: 02-300

No part of this work covered by the copyright herein may be reproduced or distributed in any form or by any means, except as permitted by U.S. copyright law, without the prior written permission of the copyright owner.

Unless otherwise noted, all content is Copyright © Cengage Learning, Inc.

Previous Edition: © 2020, © 2016, © 2012

For product information and technology assistance, contact us at
Cengage Customer & Sales Support, 1-800-354-9706
or support.cengage.com.

For permission to use material from this text or product, submit all requests online at **www.copyright.com.**

Library of Congress Control Number: 2023921539

Student Edition:
ISBN: 979-8-214-00000-8

Loose-leaf Edition:
ISBN: 979-8-214-00002-2

Cengage
5191 Natorp Boulevard
Mason, OH 45040
USA

Cengage is a leading provider of customized learning solutions.
Our employees reside in nearly 40 different countries and serve digital learners in 165 countries around the world. Find your local representative at **www.cengage.com.**

To learn more about Cengage platforms and services, register or access your online learning solution, or purchase materials for your course, visit **www.cengage.com.**

Printed in the United States of America
Print Number: 01 Print Year: 2023

To Lygia and Kyle

Roxy Peck

To my wife, Sally, and my daughter, Anna

Chris Olsen

Author Bios

Roxy Peck is Emerita Associate Dean of the College of Science and Mathematics and Professor of Statistics Emerita at California Polytechnic State University, San Luis Obispo. A faculty member at Cal Poly from 1979 until 2009, Roxy served for 6 years as Chair of the Statistics Department before becoming Associate Dean, a position she held for 13 years. She received an M.S. in Mathematics and a Ph.D. in Applied Statistics from the University of California, Riverside. Roxy is nationally known in the area of statistics education, and she was presented with the Lifetime Achievement Award in Statistics Education at the U.S. Conference on Teaching Statistics in 2009. In 2003 she received the American Statistical Association's Founder's Award, recognizing her contributions to K–12 and undergraduate statistics education. She is a Fellow of the American Statistical Association and an elected member of the International Statistics Institute. Roxy served for 5 years as the Chief Reader for the Advanced Placement (AP®) Statistics Exam and has chaired the American Statistical Association's Joint Committee with the National Council of Teachers of Mathematics on Curriculum in Statistics and Probability for Grades K–12 and the Section on Statistics Education. In addition to her texts in introductory statistics, Roxy is also co-editor of *Statistical Case Studies: A Collaboration Between Academe and Industry* and a member of the editorial board for *Statistics: A Guide to the Unknown,* 4th edition. Outside the classroom, Roxy likes to travel and spends her spare time reading mystery novels. She also collects Navajo rugs and heads to Arizona and New Mexico whenever she can find the time.

Chris Olsen taught statistics in Cedar Rapids, Iowa, for over 25 years, and at Cornell College and Grinnell College for 11 years. Chris is a past member (twice!) of the Advanced Placement® Statistics Test Development Committee and has been a table leader and question leader at the AP® Statistics reading for 16 years. He has recently retired and is spending his days and half his nights reading. Chris is a long-time consultant to the College Board, and Chris has led workshops and institutes for AP® Statistics teachers in the United States and internationally. Chris was the Iowa recipient of the Presidential Award for Excellence in Science and Mathematics Teaching in 1986, a regional awardee of the IBM Computer Teacher of the Year in 1988, and received the Siemens Award for Advanced Placement in mathematics in 1999. Chris is a frequent contributor to and moderator of the AP® Statistics Teacher Community online and is a member of the editorial board of *Teaching Statistics*. Chris graduated from Iowa State University with a major in mathematics. While acquiring graduate degrees at the University of Iowa, he concentrated on statistics, computer programming and psychometrics. In his spare time he enjoys reading and hiking. He and his wife have a daughter, Anna, who has followed in Chris' footsteps and is now teaching statistics at Grinnell College.

Brief Contents

Preface **xi**

Chapter 1 The Role of Statistics and the Data Analysis Process **1**

Chapter 2 Collecting Data Sensibly **28**

Chapter 3 Graphical Methods for Describing Data **79**

Chapter 4 Numerical Methods for Describing Data **173**

Chapter 5 Summarizing Bivariate Data **223**

Chapter 6 Probability **305**

Chapter 7 Random Variables and Probability Distributions **374**

Chapter 8 Sampling Variability and Sampling Distributions **464**

Chapter 9 Estimation Using a Single Sample **492**

Chapter 10 Hypothesis Testing Using a Single Sample **545**

Chapter 11 Comparing Two Populations or Treatments **613**

Chapter 12 The Analysis of Categorical Data and Goodness-of-Fit Tests **691**

Chapter 13 Simple Linear Regression and Correlation: Inferential Methods **728**

Chapter 14 Multiple Regression Analysis **769**

Chapter 15 Analysis of Variance **797**

Chapter 16 Nonparametric (Distribution-Free) Statistical Methods (Optional) **16-1**

Appendix: Statistical Tables **821**

Answers to Selected Odd-Numbered Exercises **841**

Index **884**

Online material, shaded in color, can be found at faculty.cengage.com.

Copyright 2025 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s). Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

Contents

Chapter 1 The Role of Statistics and the Data Analysis Process 1

1.1 Why Study Statistics? 2
1.2 The Nature and Role of Variability 3
1.3 Statistics and the Data Analysis Process 6
1.4 Types of Data and Some Simple Graphical Displays 9
Activity 1.1 Head Sizes: Understanding Variability 22
Activity 1.2 Estimating Shape Sizes 22
Activity 1.3 A Meaningful Paragraph 24
Summary: Key Concepts and Formulas 24
Chapter Review 24
Technology Notes 25

Chapter 2 **Collecting Data Sensibly** 28
2.1 Statistical Studies: Observation and Experimentation 29
2.2 Sampling 34
2.3 Simple Comparative Experiments 46
2.4 More on Experimental Design 62
2.5 Interpreting and Communicating the Results of Statistical Analyses 67
2.6 More on Observational Studies: Designing Surveys (Optional)
Activity 2.1 Facebook Friending 70
Activity 2.2 An Experiment to Test for the Stroop Effect 70
Activity 2.3 McDonald’s and the Next 100 Billion Burgers 71
Activity 2.4 Video Games and Pain Management 71
Activity 2.5 Be Careful with Random Assignment! 72
Summary: Key Concepts and Formulas 73
Chapter Review 73
Technology Notes 76

Chapter 3 **Graphical Methods for Describing Data** 79
3.1 Displaying Categorical Data: Comparative Bar Charts and Pie Charts 80
3.2 Displaying Numerical Data: Stem-and-Leaf Displays 94
3.3 Displaying Numerical Data: Frequency Distributions and Histograms 102
3.4 Displaying Bivariate Numerical Data 122
3.5 Bivariate and Multivariable Graphical Displays 131
3.6 Interpreting and Communicating the Results of Statistical Analyses 149
Activity 3.1 Locating States 158
Activity 3.2 Bean Counters! 159
Summary: Key Concepts and Formulas 160
Chapter Review 160

Online material, shaded in color, can be found at faculty.cengage.com.

Copyright 2025 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s). Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

Technology Notes 162
Cumulative Review Exercises 167

Chapter 4 **Numerical Methods for Describing Data** 173

4.1 Describing the Center of a Data Set 174
4.2 Describing Variability in a Data Set 184
4.3 Summarizing a Data Set: Boxplots 194
4.4 Interpreting Center and Variability: Chebyshev's Rule, the Empirical Rule, and z Scores 200
4.5 Interpreting and Communicating the Results of Statistical Analyses 209
Activity 4.1 Collecting and Summarizing Numerical Data 214
Activity 4.2 Airline Passenger Weights 214
Activity 4.3 Boxplot Shapes 214
Summary: Key Concepts and Formulas 215
Chapter Review 215
Technology Notes 217

Chapter 5 **Summarizing Bivariate Data** 223

5.1 Correlation 224
5.2 Linear Regression: Fitting a Line to Bivariate Data 235
5.3 Assessing the Fit of a Line 248
5.4 Nonlinear Relationships and Transformations 270
5.5 Interpreting and Communicating the Results of Statistical Analyses 288
5.6 Logistic Regression (Optional) 5-1
Activity 5.1 Age and Flexibility 294
Summary: Key Concepts and Formulas 294
Chapter Review 295
Technology Notes 298
Cumulative Review Exercises 302

Chapter 6 **Probability** 305

6.1 Chance Experiments and Events 306
6.2 Definition of Probability 313
6.3 Basic Properties of Probability 318
6.4 Conditional Probability 325
6.5 Independence 337
6.6 Some General Probability Rules 345
6.7 Estimating Probabilities Empirically and Using Simulation 357
Activity 6.1 Kisses 368
Activity 6.2 A Crisis for European Sports Fans? 369
Activity 6.3 The "Hot Hand" in Basketball 369
Summary: Key Concepts and Formulas 370
Chapter Review 370

Chapter 7 **Random Variables and Probability Distributions** 374

7.1 Random Variables 375
7.2 Probability Distributions for Discrete Random Variables 378
7.3 Probability Distributions for Continuous Random Variables 385

Online material, shaded in color, can be found at faculty.cengage.com.

7.4 Mean and Standard Deviation of a Random Variable 389
7.5 Binomial and Geometric Distributions 402
7.6 Normal Distributions 414
7.7 Checking for Normality and Normalizing Transformations 437
7.8 Using the Normal Distribution to Approximate a Discrete Distribution (Optional) 447
Activity 7.1 Is It Real? 452
Activity 7.2 Rotten Eggs? 453
Summary: Key Concepts and Formulas 453
Chapter Review 454
Technology Notes 457
Cumulative Review Exercises 460

Chapter 8 Sampling Variability and Sampling Distributions 464

8.1 Statistics and Sampling Variability 465
8.2 The Sampling Distribution of a Sample Mean 469
8.3 The Sampling Distribution of a Sample Proportion 479
Activity 8.1 Sampling Distribution of the Sample Mean Technology Activity 485
Activity 8.2 Sampling Distribution of the Sample Proportion Technology Activity 487
Activity 8.3 Do Students Who Take the SATs Multiple Times Have an Advantage in College Admissions? Technology Activity 488
Summary: Key Concepts and Formulas 490
Chapter Review 490

Chapter 9 Estimation Using a Single Sample 492

9.1 Point Estimation 493
9.2 Large-Sample Confidence Interval for a Population Proportion 498
9.3 Confidence Interval for a Population Mean 511
9.4 Interpreting and Communicating the Results of Statistical Analyses 522
9.5 Bootstrap Confidence Intervals for a Population Proportion (Optional) 527
9.6 Bootstrap Confidence Intervals for a Population Mean (Optional) 534
Activity 9.1 Getting a Feel for Confidence Level Technology Activity 538
Activity 9.2 An Alternative Confidence Interval for a Population Proportion Technology Activity 539
Activity 9.3 Verifying Signatures on a Recall Petition 540
Activity 9.4 A Meaningful Paragraph 540
Summary: Key Concepts and Formulas 541
Chapter Review 541
Technology Notes 543

Chapter 10 Hypothesis Testing Using a Single Sample 545

10.1 Hypotheses and Test Procedures 546
10.2 Errors in Hypothesis Testing 550
10.3 Large-Sample Hypothesis Tests for a Population Proportion 556
10.4 Hypothesis Tests for a Population Mean 568
10.5 Power and Probability of Type II Error 578

Copyright 2025 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s). Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

10.6 Interpreting and Communicating the Results of Statistical Analyses 587
10.7 Randomization Test and Exact Binomial Test for a Population Proportion (Optional) 590
10.8 Randomization Test for a Population Mean (Optional) 599
  Activity 10.1 Comparing the t and z Distributions Technology Activity 603
  Activity 10.2 A Meaningful Paragraph 605
  Summary: Key Concepts and Formulas 605
  Chapter Review 606
  Technology Notes 608
  Cumulative Review Exercises 610

Chapter 11 **Comparing Two Populations or Treatments** 613
11.1 Inferences Concerning the Difference Between Two Population or Treatment Means Using Independent Samples 614
11.2 Inferences Concerning the Difference Between Two Population or Treatment Means Using Paired Samples 632
11.3 Large-Sample Inferences Concerning the Difference Between Two Population or Treatment Proportions 645
11.4 Interpreting and Communicating the Results of Statistical Analyses 656
11.5 Simulation-Based Inference for Two Means (Optional) 660
11.6 Simulation-Based Inference for Two Proportions (Optional) 670
  Activity 11.1 Helium-Filled Footballs? 678
  Activity 11.2 Thinking About Data Collection 678
  Activity 11.3 A Meaningful Paragraph 678
  Summary: Key Concepts and Formulas 679
  Chapter Review 679
  Technology Notes 682

Chapter 12 **The Analysis of Categorical Data and Goodness-of-Fit Tests** 691
12.1 Chi-Square Tests for Univariate Data 692
12.2 Tests for Homogeneity and Independence in a Two-way Table 702
12.3 Interpreting and Communicating the Results of Statistical Analyses 716
  Activity 12.1 Pick a Number, Any Number 720
  Activity 12.2 Color and Perceived Taste 721
  Summary: Key Concepts and Formulas 721
  Chapter Review 722
  Technology Notes 723

Chapter 13 **Simple Linear Regression and Correlation: Inferential Methods** 728
13.1 Simple Linear Regression Model 729
13.2 Inferences About the Slope of the Population Regression Line 741
13.3 Checking Model Adequacy 753
  Activity 13.1 Sleep and Energy Intake 763
  Summary: Key Concepts and Formulas 764
  Technology Notes 764
  Cumulative Review Exercises 765

13.4 Inferences Based on the Estimated Regression Line (Optional) 13-1
13.5 Inferences About the Population Correlation Coefficient (Optional) 13-8
13.6 Interpreting and Communicating the Results of Statistical Analyses (Optional) 13-11

Chapter 14 Multiple Regression Analysis 769

14.1 Multiple Regression Models 770
14.2 Fitting a Model and Assessing Its Utility 781
Activity 14.1 Exploring the Relationship Between Number of Predictors and Sample Size 796
Summary: Key Concepts and Formulas 796
14.3 Inferences Based on an Estimated Model (Optional) 14-1
14.4 Other Issues in Multiple Regression (Optional) 14-12
14.5 Interpreting and Communicating the Results of Statistical Analyses (Optional) 14-21
Chapter Review (Optional) 14-23

Chapter 15 Analysis of Variance 797

15.1 Single-Factor ANOVA and the *F* Test 798
15.2 Multiple Comparisons 810
Activity 15.1 Exploring Single-Factor ANOVA 817
Summary: Key Concepts and Formulas 818
Technology Notes 819
15.3 The *F* Test for a Randomized Block Experiment (Optional) 15-1
15.4 Two-Factor ANOVA (Optional) 15-7
15.5 Interpreting and Communicating the Results of Statistical Analyses (Optional) 15-17
Summary: Additional Key Concepts and Formulas 15-18
Chapter Review (Optional) 15-21

Chapter 16 Nonparametric (Distribution-Free) Statistical Methods (Optional) 16-1

16.1 Distribution-Free Procedures for Inferences About a Difference Between Two Population or Treatment Means Using Independent Samples 16-2
16.2 Distribution-Free Procedures for Inferences About a Difference Between Two Population or Treatment Means Using Paired Samples 16-9
16.3 Distribution-Free ANOVA 16-19
Summary: Key Concepts and Formulas 16-26
Appendix: Tables 16-27

Appendix: Statistical Tables 821

Answers to Selected Odd-Numbered Exercises 841

Index 884

Online material, shaded in color, can be found at faculty.cengage.com.

Copyright 2025 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s). Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

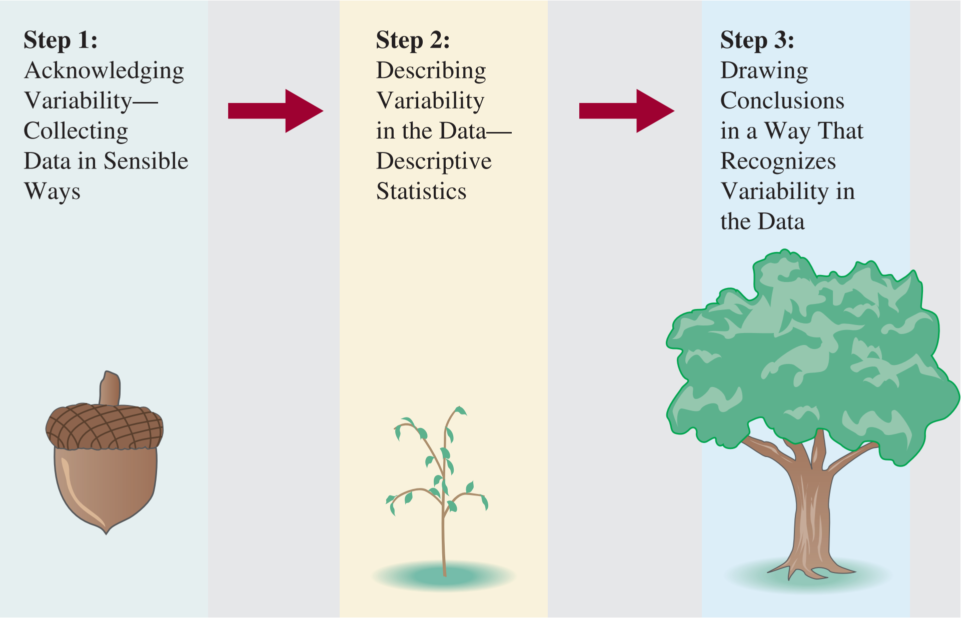
Preface

In a nutshell, statistics is about understanding the role that variability plays in drawing conclusions based on data. *Introduction to Statistics and Data Analysis,* Seventh Edition, helps introductory statistics students develop an understanding of variability through its focus on the data analysis process.

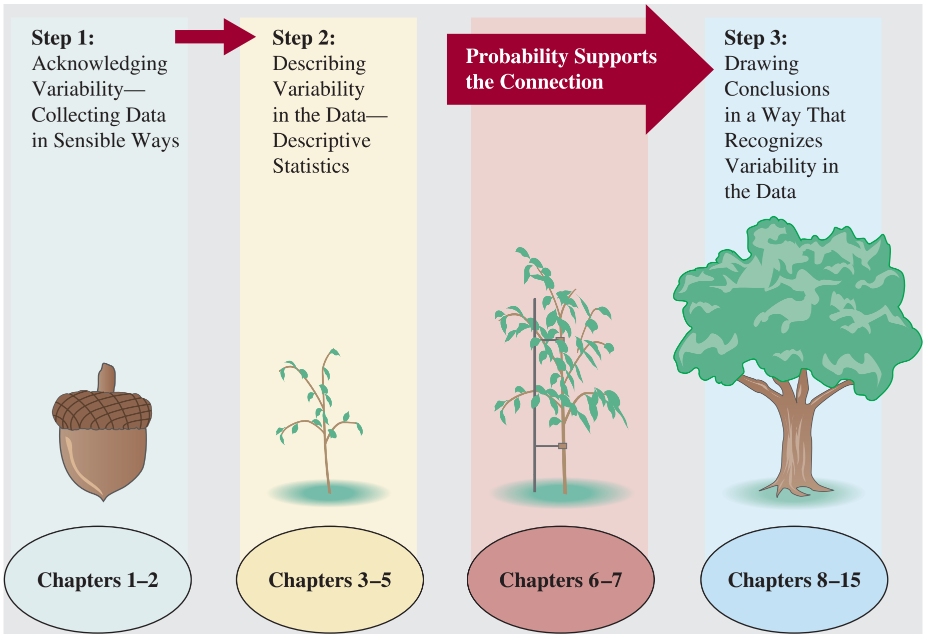
An Organization That Reflects the Data Analysis Process

Students are introduced early to the idea that data analysis is a process that begins with careful planning, followed by data collection, data description using graphical and numerical summaries, data analysis, and finally interpretation of results. This process is described in detail in Chapter 1, and the ordering of topics in the first ten chapters of the book mirrors the process: data collection, then data description, then statistical inference.

The logical order in the data analysis process can be pictured as shown in the following figure.


Step 1: Acknowledging Variability—Collecting Data in Sensible Ways
Step 2: Describing Variability in the Data—Descriptive Statistics
Step 3: Drawing Conclusions in a Way That Recognizes Variability in the Data

Unlike many introductory texts, *Introduction to Statistics and Data Analysis,* Seventh Edition, is organized in a way that is consistent with the natural order of the data analysis process:


Step 1:
Acknowledging Variability—Collecting Data in Sensible Ways
Step 2:
Describing Variability in the Data—Descriptive Statistics
Probability Supports the Connection
Step 3:
Drawing Conclusions in a Way That Recognizes Variability in the Data
Chapters 1–2
Chapters 3–5
Chapters 6–7
Chapters 8–15

The Importance of Context and Real Data

Statistics is not about numbers; it is about data—numbers in context. It is the context that makes a problem meaningful and something worth considering. For example, exercises that ask students to calculate the mean of 10 numbers or to construct a dotplot or boxplot of 20 numbers without context are arithmetic and graphing exercises. They become statistics problems only when a context gives them meaning and allows for interpretation. While this makes for a text that may appear "wordy" when compared to traditional mathematics texts, it is a critical and necessary component of a modern statistics text.

Examples and exercises with overly simple settings do not allow students to practice interpreting results in authentic situations or give students the experience necessary to be able to use statistical methods in real settings. We believe that the exercises and examples are a particular strength of this text, and we invite you to compare the examples and exercises with those in other introductory statistics texts.

Many students are skeptical of the relevance and importance of statistics. Contrived problem situations and artificial data often reinforce this skepticism. A strategy that we have employed successfully to motivate students is to present examples and exercises that involve data extracted from journal articles, newspapers, and other published sources. Most examples and exercises in the book are of this nature; they cover a very wide range of disciplines and subject areas. These include, but are not limited to, health and fitness, consumer research, psychology and aging, sports, environmental research, law and criminal justice, and entertainment.

A Focus on Interpretation and Communication

Most chapters include a section titled "Interpreting and Communicating the Results of Statistical Analyses." These sections include advice on how to best communicate the results of a statistical analysis and also consider how to interpret statistical summaries found in journals and other published sources. Subsections titled "A Word to the Wise" reminds readers of things that must be considered in order to ensure that statistical methods are employed in reasonable and appropriate ways.

Consistent with Recommendations for the Introductory Statistics Course Endorsed by the American Statistical Association

In 2005, the American Statistical Association released the report "Guidelines in Assessment and Instruction for Statistics Education (GAISE) College Report," which included the following six recommendations for the introductory statistics course:

1. Teach statistical thinking.
2. Focus on conceptual understanding.
3. Integrate real data with a context and purpose.
4. Foster active learning.
5. Use technology to explore concepts and analyze data.
6. Use assessments to improve and evaluate student learning.

In 2016, these guidelines were revised. The new guidelines reaffirmed the six recommendations and also included two new recommendations:

- Teach statistics as an investigative process of problem-solving and decision making.
- Give students experience with multivariable thinking.

Introduction to Statistics and Data Analysis, Seventh Edition, is consistent with these recommendations and supports the GAISE guidelines in the following ways:

1. **Teach statistical thinking.** Statistical thinking and statistical literacy are promoted throughout the text in the many examples and exercises that are drawn from the popular press. In addition, a focus on the role of variability, consistent use of context, and an emphasis on interpreting and communicating results in context work together to help students develop skills in statistical thinking.
2. **Focus on conceptual understanding.** Nearly all exercises are multipart and ask students to go beyond just computation. They focus on interpretation and communication, not just in the chapter sections specifically devoted to this topic, but throughout the text. The examples and explanations are designed to promote conceptual understanding. Hands-on activities in each chapter are also constructed to strengthen conceptual understanding.
3. **Integrate real data with a context and a purpose.** The examples and exercises are context driven, and the referenced sources include the popular press as well as journal articles.
4. **Foster active learning.** While this recommendation speaks more to pedagogy and classroom practice, more than 30 hands-on activities that can be used in class or assigned as homework are provided in the text and additional activities are available in the accompanying instructor resources.
5. **Use technology to explore concepts and analyze data.** The wide availability of statistical software such as Minitab, JMP, and SPSS, and online graphing capabilities has transformed both the teaching and learning of statistics. To highlight the role of technology in statistics, we have included sample output throughout the book. In addition, numerous exercises contain data that can easily be analyzed using statistical software. However, access to a particular statistical package is not assumed. Technology manuals for statistical software packages, Excel, and TI graphing calculators are available in the online materials that accompany this text. In addition, a number of activities using Shiny web apps have been incorporated into certain sections of the text to illustrate statistical concepts and demonstrate simulation-based inference methods.

   This text allows the instructor to balance the use of computers and calculators in a manner consistent with their philosophy. We don't assume the use of a particular calculator and present the calculator capabilities in a generic format.
6. **Use assessments to improve and evaluate student learning.** Assessment materials in the form of a test bank, quizzes, and chapter exams are available in the instructor

resources that accompany this text. The items in the test bank reflect the data-in-context philosophy of the text's exercises and examples.

7. **Teach Statistics as an Investigative Process of Problem-Solving and Decision Making.** A systematic approach to inference helps students to see how data are used to answer questions and to learn about the world around them. Without such a foundation, students may see the methods they are learning in their statistics course as just a loose collection of tools and may not develop a real sense of the complete data analysis process. The organization of this text helps to highlight this process, addressing methods of data collection, followed by methods for summarizing data, followed by methods for learning from data.
8. **Give Students Experience with Multivariable Thinking.** Several new examples have been included to provide students with opportunities to work with data sets that include more than two variables in order to develop multivariable thinking. New coverage of multivariable graphical displays in Chapter 3 provides additional opportunities for students to engage in multivariable thinking.

Topic Coverage

This text can be used in courses as short as one quarter or as long as one year in duration. Particularly in shorter courses, an instructor will need to be selective in deciding which topics to include and which to set aside. The book divides naturally into four major sections: collecting data and descriptive methods (Chapters 1–5), probability (Chapters 6–8), the basic one- and two-sample inferential techniques (Chapters 9–12), and more advanced inferential methods (Chapters 13–16).

We include an early chapter (Chapter 5) on descriptive methods for bivariate numerical data. This early exposure introduces questions and issues that should stimulate student interest. It is also advantageous for those teaching courses in which time constraints preclude covering advanced inferential material. However, this chapter can easily be postponed until the basics of inference have been covered, and then combined with Chapter 13 for a unified treatment of regression and correlation.

With the possible exception of Chapter 5, Chapters 1 through 10 should be covered in order. We anticipate that most instructors will then continue with two-sample inference (Chapter 11) and methods for categorical data analysis (Chapter 12), although regression could be covered before either of these topics. Optional portions of Chapter 14 (multiple regression) and Chapter 15 (analysis of variance) and Chapter 16 (nonparametric methods) are included in the online materials that accompany this text.

A Note on Probability

The content of the probability chapters include both a traditional treatment of probability and probability distributions at an introductory level, as well as a section on the use of simulation as a tool for estimating probabilities. For those who prefer a more informal treatment of probability, see the text *Statistics: Learning from Data*, Third Edition, by Roxy Peck and Catherine Case.

In This Edition

Look for the following in the Seventh Edition:

- **NEW Updated Examples and Exercises.** In our continuing effort to keep things interesting and relevant, this edition contains many updated examples and exercises that use data from recent journal articles, newspapers, and web posts, on topics of interest to students.
- **NEW Section on Bivariate and Multivariable Graphical Displays.** Graphical displays and data visualizations are increasingly common ways to summarize data and

Copyright 2025 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s). Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

communicate findings. Many of the displays that students will encounter in the media will contain information on more than one or two variables and are more complex than traditional graphs like histograms, boxplots, and bar charts. A new section has been added to Chapter 3 (Section 3.5) that looks at several additional types of graphical displays and provides insight into what students should look for when they encounter more complex or interactive graphical displays in the media. This section is designed to help students think critically when interpreting multivariable graphical displays.

- **NEW What Do You Think? boxes.** To help strengthen students' conceptual understanding, these questions ask students to pause periodically to apply what they've learned to a particular situation.
- **Sections on Randomization-Based Inference Methods.** Research indicates that randomization-based instruction in statistical inference may help learners to better understand the concepts of confidence and significance. This edition includes optional sections on randomization-based inference methods. These methods are also particularly useful in that they provide an alternative method of analysis that can be used when the conditions required for normal distribution-based inference are not met. The inference chapters (Chapters 9–11) contain optional sections on randomization-based inference that include bootstrap methods for simulation-based confidence intervals and randomization-based tests of hypotheses. These sections are accompanied by online Shiny apps, which can be used to construct bootstrap confidence intervals and to carry out randomization tests.
- **Helpful hints in exercises.** To help students who might be having trouble getting started, hints are provided with many of the exercises directing students to relevant examples in the text.
- **Margin notes to highlight the importance of context and the process of data analysis.** Margin notations appear in appropriate places in the examples. These include Understanding the context, Consider the data, Formulate a plan, Do the work, and Interpret the results. These notes are designed to increase student awareness of the steps in the data analysis process.
- **Activities at the end of each chapter.** These activities can be used as a chapter capstone or can be integrated at appropriate places as the chapter material is covered in class.
- **Advanced Topics.** Topics that are often omitted in a one-quarter or one-semester course, such as survey design (Section 2.6), logistic regression (Section 5.6), inference based on the estimated regression line (Sections 13.4 and 13.5), inference and variable selection methods in multiple regression (Sections 14.3 and 14.4), analysis of variance for randomized block and two-factor designs (Sections 15.3 and 15.4), and distribution-free procedures (Chapter 16) are available online.

Instructor and Student Resources

WebAssign

Built by educators, WebAssign provides flexible settings at every step to customize your course with online activities to meet learners' unique needs. Students get everything in one place, including rich content and study resources designed to fuel deeper understanding, plus access to a dynamic, interactive eTextbook. The engaging variety of WebAssign problem types gain student interest, including Select Your Scenario Questions, Stats in Practice Questions, and Project Milestones. Students can explore real data with questions linked to the Statistical Analysis & Learning Tool (SALT). You can also utilize the Intro Stats Readiness Gaps to get students up to speed with prerequisites and the Statistics Question Collection to provide additional practice throughout the term. For more information, visit https://www.cengage.com/webassign.

Copyright 2025 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s). Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

Statistical Analysis & Learning Tool (SALT)

The Statistical Analysis & Learning Tool—known as SALT—enables Introductory Statistics students to analyze and visualize data and think statistically without getting bogged down in complex computations. Within WebAssign, you can assign exercises in which students use SALT to analyze data and correctly answer the questions. The ease-of-use enables students to spend time learning statistics, not how to use a data analysis tool.

Applets

A collection of easy to use applets designed to support simulation-based inference is available at https://www.cengage.com/Peck. Included are applets that support the sections on bootstrap confidence intervals and randomization tests, as well as applets that help students visualize the meaning of confidence level and understand the concept of sampling variability.

Instructor Resources

Additional resources for this product are available in the Cengage Instructor Center. Instructor assets include a Complete Solutions Manual, PowerPoint® slides, Guide to Teaching Online, Educator's Guide, a test bank powered by Cognero®, and more. Sign up or sign in at faculty.cengage.com to search for and access this product and its online resources.

- Complete Solutions Manual—provides solutions and answers to textbook questions.
- PowerPoint® slides—support lectures with definitions, formulas, examples, and activities.
- Instructor Manual—outlines activities and resources that can be incorporated into instruction for each chapter.
- Guide to Teaching Online—offers tips for teaching online and incorporating WebAssign activities into your course.
- Educator's Guide—offers suggested content for WebAssign by chapter to help you personalize your course.
- Cengage Testing, powered by Cognero®—a flexible, online system that allows you to access, customize, and deliver a test from your chosen text's test bank to your students through your LMS or another channel outside of WebAssign.
- Data Sets—provide the data from textbook examples and exercises in downloadable files.
- Technology Guides—help students work through problems using Excel and R.
- Transition Guide—outlines changes between the 6th and 7th editions of the textbook.

Acknowledgments

We are grateful for the thoughtful feedback from the following reviewers that has helped to shape this text over the previous editions:

Reviewers for the Seventh Edition

Nimet Alpay, Franklin University

Paul Holmes, University of Georgia

Shahera Jannah, CUNY

Fidele Ngwane, University of South Carolina, Salkehatchie

Alex Rolón, Northampton Community College

Reviewers for Previous Editions

Arun K. Agarwal, Jacob Amidon, Holly Ashton, Barb Barnet, Chad Bemis, Eddie Bevilacqua, Piotr Bialas, Kelly Black, Jim Bohan, Carl Brezovec, Pat Buchanan, Gabriel Chandler, Andy Chang, Jerry Chen, Richard Chilcoat, Mary Christman, Charles Conrad, Christina Cornejo, Marvin Creech, Kathleen Dale, Ron Degged, Hemangini Deshmukh, Ann Evans, Guangxiong Fang, Sharon B. Finger, Donna Flint, Steven Garren, Mark Glickman, Rick Gumina, Debra Hall, Tyler Haynes, Sonja Hensler, Zhongming Huang, Trish Hutchinson, John Imbrie, Bambi Jones, Bessie Kirkwood, Jeff Kollath, Christopher Lacke, Austin Lampros, Michael Leitner, Zia Mahmood, David Manley, Art Mark, Pam Martin, David Mathiason, Bob Mattson, Kendra Mhoon, C. Mark Miller, Megan Mocko, Paul Myers, Kane Nashimoto, Helen Noble, Douglas Noe, Broderick Oluyede, Elaine Paris, Shelly Ray Parsons, Deanna Payton, Judy Pennington-Price, Greg Perkins, Michael Phelan, Alan Polansky, John Racquet, Mamunur Rashid, Leah Rathbun, Michael Ratliff, David Rauth, Kevin J. Reeves, Lawrence D. Ries, Hazel Shedd, Robb Sinn, Greg Sliwa, Angela Stabley, Jeffery D. Sykes, Weizhong Tian, Yolanda Tra, Joe Ward, Nathan Wetzel, Mark Wilson, Yong Yu, and Toshiyuki Yuasa, Cathleen Zucco-Teveloff.

We would also like to express our thanks and gratitude to those whose support made this seventh edition possible:

- Adrian Daniel, Associate Portfolio Product Manager
- Lynh Pham, Content Manager
- Elinor Gregory, Senior Learning Designer
- Prashant Das, our project manager at MPS Limited
- Stephen Miller for his work in creating new student and instructor solutions manuals to accompany the text
- Kellie Woodson for checking the accuracy of examples and solutions

And, as always, we thank our families, friends, and colleagues for their continued support.

Roxy Peck
Chris Olsen

AP® is a trademark registered by the College Board, which is not affiliated with, and does not endorse this product.

Introduction to Statistics and Data Analysis

Chapter

1

The Role of Statistics and the Data Analysis Process

Studio Romantic/Shutterstock.com

Statistics helps you to make sense of data and to gain insight into the world around you. The use of statistical analyses in fields such as business, medicine, agriculture, social science, natural science, and engineering makes statistical literacy an important part of a well-rounded education.

Statistical methods help you to make intelligent judgments and informed decisions in the presence of uncertainty and variability. In this chapter, the role of variability in statistical settings and some simple graphical displays for summarizing data are considered.

Learning Objectives

After completing this chapter, you will be able to:

- Describe the steps in the data analysis process.
- Distinguish between a population and a sample.
- Distinguish between categorical, discrete numerical, and continuous numerical data.
- Construct frequency distributions and bar charts using categorical data, and dotplots using numerical data.
- Describe the distribution of a categorical or numerical variable.

Section 1.1 Why Study Statistics?

There is an old saying that "without data, you are just another person with an opinion." While anecdotes and coincidences may make for interesting stories, you wouldn't want to make important decisions on the basis of anecdotes alone. For example, just because a friend of a friend ate 16 apricots and then experienced relief from joint pain doesn't mean that this is all you need to know to help one of your parents choose a treatment for arthritis! Before recommending apricots, you would definitely want to consider relevant data—that is, data that would allow you to investigate the effectiveness of apricots as a treatment for arthritis.

It is challenging to function in today's world without a basic understanding of statistics. For example, here are just a few conclusions based on data from articles that appeared in a single newspaper during a 2-week period in December, 2022.

Electric vehicles, smartphones, and laptops all use lithium-ion batteries, which contain reusable materials such as cobalt, lithium, and nickel. These materials are expensive and in short supply, so with the increased demand for lithium-ion batteries, some companies are looking at ways to recycle and reclaim these materials. The article **"Electric Cars Are Taking Off, but When Will Battery Recycling Follow?" (*The New York Times*, December 21, 2022)** looked at the current state of recycling of cobalt and lithium and provided an estimate of the percentages of recycling that will consist of scraps from manufacturing and from batteries by 2025.

"Betting Apps Can Make Anyone a Sports Fan. Even Me" (*The New York Times*, December 12, 2022) is the title of an article that looked at the growth of online sports betting since 2018, when it became legal in some states. Based on data from the Pew Research Center, the article reported that about 20% of adult Americans had bet on sports in the past year and that in the first half of 2022, about $8 billion per month in legal sports bets were placed. Some projections of the dollar value of legal sports bets in 2026 were reported to be as high as $20 billion per month.

Housing prices are high in many large cities in the United States. The Department of Housing and Urban Development recommends that no more than 30% of monthly income be spent on housing. Is this attainable? The article **"Where Are the Most and Least Affordable Homes?" (The *New York Times*, December 15, 2022)** described an analysis of home prices and household incomes to rank the 100 largest U.S. cities based on housing affordability. The least affordable cities were Miami, Los Angeles, and New York. The most affordable cities were Wichita, Fort Wayne, and Detroit. The article included a table listing the median home price and median income for the 10 most and the 10 least affordable cities. (In a data set, the median is the number that has half of the data values above and half below.)

The article **"Despite Modest G.O.P. Gains, Democrats Maintain Grasp on Suburbs" (*The New York Times*, December 13, 2022)** provided an analysis of data from the 2016, 2020, and 2022 elections to reach the conclusion that Democrats gained ground in the suburbs of large cities in the 2020 elections compared to 2016. They also concluded that while Republicans recovered in these areas in the 2022 election, the gain was smaller than the previous shift toward Democrats. These conclusions were supported by a graphical display showing changes in voting patterns from 2016 to 2020 and from 2020 to 2022 for the 25 largest U.S. cities.

"It's Boom Times for Santa" (*The New York Times*, December 23, 2022) is the claim made in the title of a light-hearted article that compared the wage growth from 2015 to 2022 for five types of in-person jobs and for all workers. The article notes that the number of bookings for Santas were up during the holiday season in 2022 and that wages earned by Santas also increased dramatically between 2020 and 2022. A graph in the article showed that Santa wages increased over 40% between 2015 and 2022, and showed a drop in 2020 followed by steep increase. The change over time for Santa wages outpaced those working in other categories that were also shown in the graph, including leisure and hospitality, retail, and hospital workers.

To be an informed consumer of reports such as those described above, you must be able to do the following:

1. Extract information from tables, charts, and graphs.
2. Follow numerical arguments.
3. Understand the basics of how data should be gathered, summarized, and analyzed in order to draw statistical conclusions.

Your statistics course will help prepare you to perform these tasks.

Studying statistics will also enable you to collect data in a sensible way and then use the data to answer questions of interest. In addition, studying statistics will allow you to critically evaluate the work of others by providing you with the tools you need to make informed judgments.

Throughout your personal and professional life, you will need to understand and use data to make decisions. To do this, you must be able to

1. Decide whether existing data are adequate or whether additional information is required.
2. If necessary, collect more information in a reasonable and thoughtful way.
3. Summarize available data in a useful and informative manner.
4. Analyze available data.
5. Draw conclusions, make decisions, and assess the risk of an incorrect decision.

These are the steps in the data analysis process. These steps are considered in more detail in Section 1.3.

We hope that this textbook will help you to understand the logic behind statistical reasoning, prepare you to apply statistical methods appropriately, and enable you to recognize when statistical arguments are faulty.

Section 1.2 The Nature and Role of Variability

Statistical methods allow you to collect, describe, analyze, and draw conclusions from data. If you lived in a world where all measurements were identical for every individual, these tasks would be simple. Imagine a population consisting of all students at a particular university. Suppose that *every* student is enrolled in the same number of courses, spent exactly the same amount of money on textbooks this semester, and favors increasing student fees to support expanding library services. For this population, there is *no* variability in number of courses, amount spent on books, or student opinion on the fee increase. A researcher studying students from this population in order to draw conclusions about these three variables would have a particularly easy task. It would not matter how many students the researcher studied or how the students were selected. In fact, the researcher could collect information on number of courses, amount spent on books, and opinion on the fee increase by just stopping the next student who happened to walk by. Because there is no variability in the population, this one individual would provide complete and accurate information about the population. The researcher could draw conclusions with no risk of error.

The situation just described is obviously unrealistic. Populations with no variability are rare. You need to understand variability to be able to collect, describe, analyze, and draw conclusions from data in a sensible way.

Examples 1.1 and 1.2 illustrate how describing and understanding variability are important.

Example 1.1 If the Shoe Fits

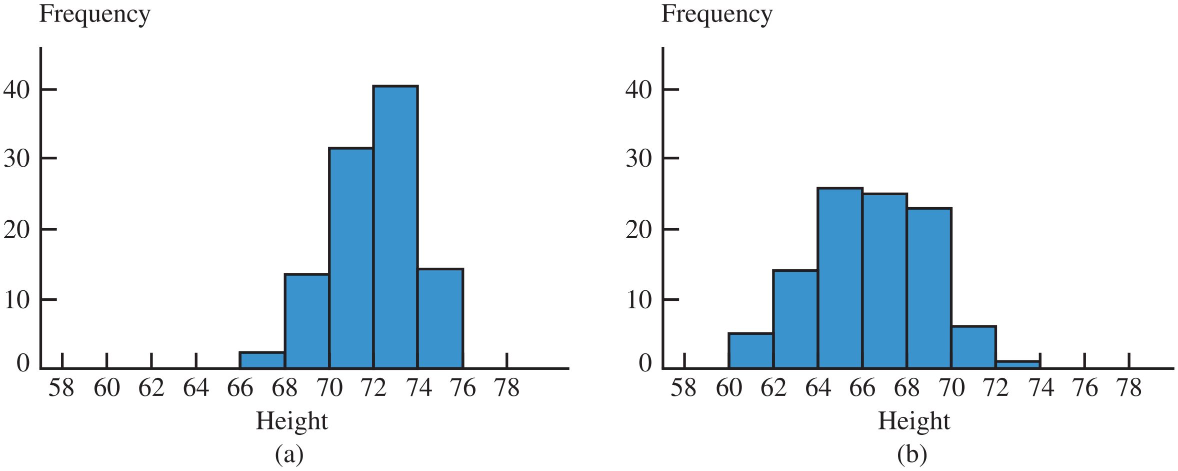
Understand the context ➤

The graphs in Figure 1.1 are examples of a type of graph called a histogram. (The construction and interpretation of histograms is discussed in Chapter 3.) Figure 1.1(a) shows the distribution of the heights of women's basketball players who played at a particular university. Each bar in the histogram represents a particular range of player heights. The height of each bar in the graph indicates how many players' heights were in the corresponding range. For example, 40 basketball players had heights between 72 inches and 74 inches, while only 2 players had heights between 66 inches and 68 inches. Figure 1.1(b) shows the distribution of heights for members of the women's gymnastics team. Both histograms are based on the heights of 100 athletes.

Consider the data ➤

Figure 1.1

Histograms of heights (in inches): (a) 100 women's basketball players; (b) 100 women's gymnastics team members.



The first histogram shows that the heights of women's basketball players varied, with most heights falling between 68 inches and 76 inches. In the second histogram you can see that the heights of women's gymnastics team members also varied, with most heights in the range of 60 inches to 72 inches. It is also clear that there is more variation in the heights of the gymnasts than in the heights of the basketball players, because the gymnast histogram spreads out more about its center than does the basketball histogram.

Now suppose that a tall woman (5 feet 11 inches) tells you she is looking for her sister who is practicing with her team at the gym.

What Do You Think?

Should you direct her to where the basketball team is practicing or to where the gymnastics team is practicing? What reasoning did you use to decide? If you found a pair of size 6 shoes left in the locker room, should you first try to return them by checking with members of the basketball team or the gymnastics team?

Interpret the results ➤

You would probably send the woman looking for her sister to the basketball practice and you would probably try to return the shoes to a gymnastics team member. Reaching these conclusions requires statistical reasoning that combines knowledge of the relationship between heights of siblings and between shoe size and height with the information about the distributions of heights presented in Figure 1.1. You might have reasoned that heights of siblings tend to be similar and that a height as great as 5 feet 11 inches, although not impossible, would be unusual for a gymnast. On the other hand, a height as tall as 5 feet 11 inches would be common for a basketball player.

Copyright 2025 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s). Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

Similarly, you might have reasoned that tall people tend to have bigger feet and that short people tend to have smaller feet. The shoes found were a small size, so it is more likely that they belong to a gymnast than to a basketball player, because small heights are common for gymnasts and unusual for basketball players.

Example 1.2 Monitoring Water Quality

Understand the context ➤

As part of its regular water quality monitoring efforts, an environmental control board selects five containers of water from a particular well each day. The concentration of contaminants in parts per million (ppm) is measured for each of the five containers, and then the average of the five measurements is calculated. The histogram in Figure 1.2 summarizes the average contamination values for 200 days.

Now suppose that a chemical spill has occurred at a manufacturing plant 1 mile from the well. It is not known whether a spill of this nature would contaminate groundwater in the area of the spill and, if so, whether a spill this distance from the well would affect the quality of well water.

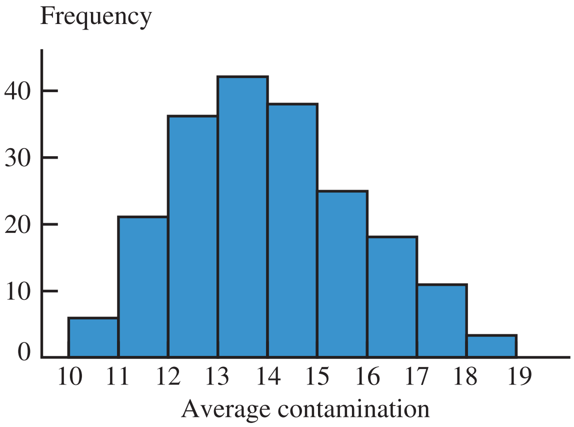
Consider the data ➤

One month after the spill, five containers of water are collected from the well, and the average contamination is 15.5 ppm.

What Do You Think?

Considering the variability before the spill, should you interpret this as evidence that the well water was affected by the spill? What if the calculated average was 17.4 ppm? 22.0 ppm? How is your reasoning related to the histogram in Figure 1.2?

Figure 1.2
Average contamination concentration (in ppm) measured each day for 200 days.



Interpret the results ➤

Before the spill, the average contaminant concentration varied from day to day. An average of 15.5 ppm would not have been an unusual value, so seeing an average of 15.5 ppm after the spill isn't necessarily an indication that contamination has increased. On the other hand, an average as large as 17.4 ppm is less common, and an average as large as 22.0 ppm is not at all typical of the pre-spill values. In this case, you would probably conclude that the well contamination level has increased.

In these two examples, reaching a conclusion required an understanding of variability. Understanding variability allows you to distinguish between common and unusual values. The ability to recognize unusual values in the presence of variability is an important aspect of most statistical procedures. It also enables you to quantify the chance of being incorrect when a conclusion is based on data. These concepts are developed further in later chapters.

Section 1.3 Statistics and the Data Analysis Process

Statistical studies are undertaken to answer questions about the world. Is a new flu vaccine effective in preventing illness? Is the use of bicycle helmets on the rise? Are injuries that result from bicycle accidents less severe for riders who wear helmets than for those who do not? Data collection and analysis allow researchers to answer questions like these.

The data analysis process can be viewed as a sequence of steps that lead from planning to data collection to making informed conclusions based on the resulting data. The process can be organized into the six steps described below.

The Data Analysis Process

1. **Understanding the nature of the problem.** Effective data analysis requires an understanding of the research problem. You must know the goal of the research and what questions you hope to answer. It is important to have a clear direction before gathering data to ensure that you will be able to answer the questions of interest using the data collected.
2. **Deciding what to measure and how to measure it.** The next step in the process is deciding what information is needed to answer the questions of interest. In some cases, the choice is obvious. For example, in a study of the relationship between the weight of a Division I football player and position played, you would need to collect data on player weight and position. In other cases the choice of information is not as straightforward. For example, in a study of the relationship between preferred learning style and intelligence, how should you define learning style and measure it? What measure of intelligence should you use? It is important to carefully define the variables to be studied and develop appropriate methods for determining their values.
3. **Data collection.** The data collection step is very important. The researcher must first decide whether an existing data source is adequate or whether new data must be collected. If a decision is made to use existing data, it is important to understand how the data were collected and for what purpose, so that any resulting limitations are also fully understood. If new data are to be collected, a careful plan must be developed, because the type of analysis that is appropriate and the conclusions that can be drawn depend on how the data are collected.
4. **Data summarization and preliminary analysis.** After the data are collected, the next step is usually a preliminary analysis that includes summarizing the data graphically and numerically. This initial analysis provides insight into important characteristics of the data and provides guidance in selecting appropriate methods for further analysis.
5. **Formal data analysis.** The data analysis step requires the researcher to select appropriate statistical methods. Much of this textbook is devoted to methods that can be used to carry out this step.
6. **Interpretation of results.** Several questions should be addressed in this final step. Some examples are: What can be learned from the data? What conclusions can be drawn from the analysis? How can the results guide future research? The interpretation step often leads to the formulation of new research questions. These new questions lead back to the first step. In this way, good data analysis is often an iterative process.

To illustrate these steps, consider the following example. The admissions director at a large university might be interested in learning why some applicants who were accepted for the fall semester failed to enroll at the university. The **population** of interest to the director consists of all accepted applicants who did not enroll in the fall semester. Because this population is large and it may be difficult to contact all the individuals, the director might decide to collect data from only 300 of these students. These 300 constitute a **sample**.

Definitions

Population: The entire collection of individuals or objects about which information is desired is called the **population** of interest.

Sample: A **sample** is a subset of the population.

Deciding how to select the 300 students and what data should be collected from each student are steps 2 and 3 in the data analysis process. Step 4 in the process involves organizing and summarizing data. Methods for organizing and summarizing data, such as the use of tables, graphs, and numerical summaries, make up the branch of statistics called **descriptive statistics**. A second major branch of statistics, **inferential statistics**, involves generalizing from a sample to the population from which it was selected. When you generalize in this way, you run the risk of an incorrect conclusion, because a conclusion about the population is based on incomplete information. Because of this, it is important to look at ways to quantify the chance of an incorrect conclusion.

Definitions

Descriptive statistics: The branch of statistics that includes methods for organizing and summarizing data.

Inferential statistics: The branch of statistics that involves generalizing from a sample to the population from which the sample was selected and assessing the reliability of such generalizations.

Example 1.3 illustrates the steps in the data analysis process.

Example 1.3 Chew More, Eat Less?

Understand the context ➤

The article **"Increasing the Number of Chews before Swallowing Reduces Meal Size in Normal-Weight, Overweight, and Obese Adults" (*Journal of the Academy of Nutrition and Dietetics* [2014]: 926–931)** describes a study that investigated whether chewing each bite of food more before swallowing would result in people eating less. Participants in the study were adults between the ages of 18 and 45 years. At the beginning of the study, each participant was observed as they each ate five pizza rolls, and the number of chews made before swallowing was observed in order to determine a baseline for that participant.

Participants were then invited back for a second session on a different day. They were asked to eat their usual breakfast on that day and to not eat anything after breakfast. At the second session, all participants were provided with a platter of pizza rolls and were told to eat until they were comfortably full. They were also told they could request more pizza rolls if they wanted more. Each participant was also told how many times to chew each pizza roll before swallowing. Then, each participant was assigned to one of three groups. The participants in group 1 were given a number of chews equal to their baseline. The participants in group 2 were given a number of chews that was 150% of (one and a half times as large as) their baseline. The participants in group 3 were assigned a number of chews that was 200% of (twice as large as) their baseline.

Interpret the results ➤

After analyzing data from this study, the researcher concluded that people ate about 10% less when they increased the number of chews by 50% (group 2) and about 15% less when they doubled the number of chews.

This study illustrates the nature of the data analysis process. A clearly defined research question and an appropriate choice of how to measure the variables of interest (the number of chews and how much people ate) preceded the data collection. Assuming that a reasonable method was used to collect the data (you will see how this can be evaluated in Chapter 2) and that appropriate methods of analysis were employed, the investigators reached the conclusion that increasing the number of chews before swallowing results in people tending to eating less.

Exercises 1.1–1.11

1.1 Give brief definitions of the terms *descriptive statistics* and *inferential statistics.*

1.2 Give brief definitions of the terms *population* and *sample.*

1.3 The **National Highway Traffic Safety Administration (crashstats.nhtsa.dot.gov)** reported that in 2018, 857 cyclists were killed on public roadways in the United States and that the average age of the cyclists killed was 47 years. These figures were based on an analysis of the records of all traffic related deaths of cyclists on U.S. public roadways in 2018.

- **a.** Does the group of 857 cycling fatalities represent a census or a sample of the cycling fatalities in 2018?
- **b.** If the population of interest is cycling fatalities in 2018, is the given average age of 47 years a population value or was it calculated from a sample?

1.4 **"San Fernando Valley Residents OK with 1-Cent Transit Tax, MTA Poll Says"** is the headline of an article that appeared in the ***LA Daily News* (April 12, 2016)**. This headline was based on responses from a sample of 100 San Fernando Valley residents. Describe the sample and the population of interest for this study.

1.5 The student senate at a university with 15,000 students is interested in the proportion of students who favor a change in the grading system to allow for plus and minus grades (for example, B+, B, B−, rather than just B). Two hundred students are interviewed to determine their attitude toward this proposed change.

- **a.** What is the population of interest?
- **b.** What group of students constitutes the sample in this problem?

1.6 Data from a survey of 10,413 students and 588 teachers conducted for the Knight Foundation were used to calculate the following estimates: 65% of students and 40% of teachers chose freedom of speech as the most important of the rights guaranteed by the First Amendment **(*USA TODAY*, October 3, 2014)**. Are the percentages given calculated using data from a sample of for the entire population?

1.7 The supervisors of a rural county are interested in the proportion of property owners who support the construction of a sewer system. Because it is too costly to contact all 7000 property owners, a survey of 500 owners is undertaken. Describe the population and sample for this problem.

1.8 A consumer group conducts crash tests of new model cars. To determine the severity of damage to 2023 Toyota Camrys resulting from a 10-mph crash into a concrete wall, the research group tests six cars of this type and assesses the amount of damage. Describe the population and the sample for this problem.

1.9 A building contractor has a chance to buy an odd lot of 5000 used bricks at an auction. They are interested in determining the proportion of bricks in the lot that are cracked and therefore unusable for their current project, but they do not have enough time to inspect all 5000 bricks. Instead, they check 100 bricks to determine which ones are cracked. Describe the population and the sample for this problem.

1.10 The authors of the article **"Can Artificial Intelligence Pass the Fellowship of the Royal College of Radiologists Examination?" (*British Medical Journal* [2022]: e072826)** wondered if an artificial intelligence (AI) program designed to interpret radiographs (medical images such as X-rays and MRI scans) was as accurate as trained radiologists. To investigate, they created 10 mock exams that were similar to the ones that radiologists in England are required to pass before they can practice independently. They found that the AI program passed only 2 of the 10 exams and was unable to interpret a number of the images on the exams.

- **a.** What were the researchers trying to learn? What questions motivated their research?
- **b.** Do you think that the study was conducted in a reasonable way? What additional information would you want in order to evaluate this study?

1.11 In a study to investigate the effect of blue light exposure on the quality and ripeness of mangoes, researchers placed a sample of mangoes under blue light for 9 days and placed another sample of mangoes in darkness for 9 days **("Blue Light Simultaneously Induces Peel Anthocyanin Biosynthesis and Flesh Carotenoid/Sucrose Biosynthesis in Mango Fruit," *Journal of Agricultural and Food Chemistry* [2022]: 16021–16035)**. At the end of the 9 days, they measured ripeness and sugar level of the mangoes. Based on the data from the study, the researchers concluded that the mangoes exposed to blue light were riper and sweeter than the mangoes that were kept in the dark.

- **a.** What were the researchers trying to learn? What questions motivated their research?
- **b.** Do you think that the study was conducted in a reasonable way? What additional information would you want in order to evaluate this study?

Copyright 2025 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s). Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

Section 1.4 Types of Data and Some Simple Graphical Displays

Every discipline has its own particular way of using common words, and statistics is no exception. You will recognize some of the terminology from previous math and science courses, but much of the language of statistics will be new to you. In this section, you will learn some of the terminology used to describe data. While most types of graphical displays are covered in Chapter 3, two simple graph types are introduced here because you will need to use them in Chapter 2.

Types of Data

The individuals or objects in any particular population might possess many characteristics that could be studied. Consider a group of students currently enrolled in a statistics class. One characteristic of the students in the population is the brand of calculator they use (Casio, Hewlett-Packard, Sharp, Texas Instruments, and so on). Another characteristic is the number of textbooks purchased that semester, and another is the distance from the college to a student's home. A **variable** is any characteristic whose value may change from one individual or object to another. For example, *calculator brand* is a variable, and so are *number of textbooks purchased* and *distance to the college*. **Data** result from making observations either on a single variable or on two or more variables at the same time.

Definitions

Variable: A characteristic whose value may change from one observation to another.

Data: A collection of observations on one or more variables.

A **univariate data set** consists of observations on a single variable made on individuals in a sample or population. There are two types of univariate data sets: **categorical** (sometimes also called qualitative) and **numerical** (sometimes also called quantitative). In the previous example, *calculator brand* is a categorical variable, because each student's response to the query, "What brand of calculator do you use?" is a category. The collection of responses from all these students forms a categorical data set. The other two variables, *number of textbooks purchased* and *distance to the college*, are both numerical in nature. Determining the values of a numerical variable (by counting or measuring) results in a numerical data set.

Definitions

Univariate data set: A data set consisting of observations on a single characteristic.

Categorical data set: A univariate data set is **categorical** (or **qualitative**) if the individual observations are categorical responses.

Numerical data set: A univariate data set is **numerical** (or **quantitative**) if each observation is a number.

Example 1.4 College Choice Do-Over?

Understand the context ➤

The Higher Education Research Institute at UCLA surveys over 20,000 college seniors each year. One question on the survey asks seniors the following question: If you could make your college choice over, would you still choose to enroll at your current college? Possible responses are definitely yes (DY), probably yes (PY), probably no (PN), and definitely no (DN). Responses for 20 students were:

DY	PN	DN	DY	PY	PY	PN	PY	PY	DY
DY	PY	DY	DY	PY	PY	DY	DY	PN	DY

(These data are just a small subset of the data from the survey.) Because the response to the question about college choice is categorical, this is a univariate categorical data set.

In Example 1.4, the data set consists of observations on a single variable (college choice response), so this is a univariate data set. In some studies, we are interested in two different characteristics. For example, both height (in inches) and weight (in pounds) might be recorded for each person on a basketball team. The resulting data set consists of pairs of numbers, such as (74, 185). This is called a **bivariate data set**. **Multivariate data** result from recording a value for each of two or more attributes (so bivariate data are a special case of multivariate data). For example, multivariate data would result from determining height, weight, pulse rate, and blood pressure for each person on a basketball team.

What Do You Think?

What is an example of a bivariate data set that includes one categorical variable and one numerical variable?

What is an example of a multivariate data set?

Example 1.5 How Safe Are College Campuses?

Understand the context ➤

● Consider the accompanying data on violent crime on public college campuses in Florida during 2016 **(fbi.gov)**.

Consider the data ➤

University/College	Student Enrollment	Number of Violent Crimes Reported in 2016
Florida A&M University	9,928	6
Florida Atlantic University	30,380	6
Florida Gulf Coast University	14,833	3
Florida International University	49,782	15
Florida South Western State College	15,709	0
Florida State University, Tallahassee	40,830	20
New College of Florida	861	1
Pensacola State College	9,840	0
Santa Fe College	14,767	0
Tallahassee Community College	12,445	2
University of Central Florida	62,953	14
University of Florida	50,645	10
University of North Florida	15,675	3
University of South Florida, St. Petersburg	4,739	0
University of South Florida, Tampa	42,067	8
University of West Florida	12,763	10

Here two variables—*student enrollment* and *number of violent crimes reported*—were recorded for each of the 16 schools. Because this data set consists of values of two variables for each school, it is a bivariate data set. Each of the two variables considered here is numerical (rather than categorical).

Two Types of Numerical Data

There are two different types of numerical data: **discrete** and **continuous**. Consider a number line (Figure 1.3) for locating values of a numerical variable. Each possible number (2, 3.125, 8.12976, etc.) corresponds to exactly one point on the number line.

● **Data set available online**

Copyright 2025 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s). Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

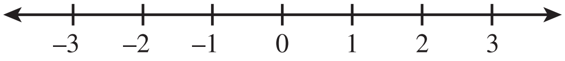
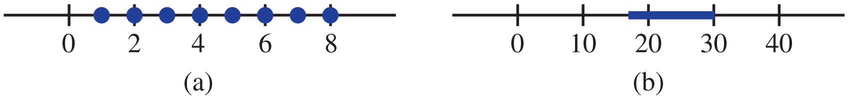


Figure 1.3
A number line.

Suppose that the variable of interest is the number of courses in which a student is enrolled. If no student is enrolled in more than eight courses, the possible values are 1, 2, 3, 4, 5, 6, 7, and 8. These values are identified in Figure 1.4(a) by the dots at the points marked 1, 2, 3, 4, 5, 6, 7, and 8. These possible values are isolated from one another on the number line. You can place an interval around any possible value that is small enough that no other possible value is included in the interval. On the other hand, the line segment in Figure 1.4(b) identifies a plausible set of possible values for the time (in seconds) it takes for the first kernel in a bag of microwave popcorn to pop. Here the possible values make up an entire interval on the number line, and no possible value is isolated from other possible values.



(a) (b)

Figure 1.4
Possible values of a variable:
(a) number of courses;
(b) popping time (in seconds).

Definitions

Discrete numerical variable: A numerical variable results in **discrete** data if the possible values of the variable correspond to isolated points on the number line.

Continuous numerical variable: A numerical variable results in **continuous** data if the set of possible values forms an entire interval on the number line.

Discrete data usually arise when observations are determined by counting (for example, the number of roommates a student has or the number of petals on a flower).

Example 1.6 **Do U Txt?**

● The number of text messages sent on a particular day is recorded for each of 12 students. The resulting data set is

23 0 14 13 15 0 60 82 0 40 41 22

Possible values for the variable *number of text messages sent* are 0, 1, 2, 3,… . These are isolated points on the number line, so this data set consists of discrete numerical data.

Suppose that instead of the number of text messages sent, the *time spent texting* had been recorded. Even though time spent may have been reported rounded to the nearest minute, the actual time spent could have been 6 minutes, 6.2 minutes, 6.28 minutes, or any other value in an entire interval. So, recording values of *time spent texting* would result in continuous data.

In general, data are continuous when observations involve making measurements, as opposed to counting. In practice, measuring instruments do not have infinite accuracy, so possible measured values, strictly speaking, may not form an entire interval on the number line. However, any number in the interval *could* be a value of the variable. The distinction between discrete and continuous data will be important in the discussion of probability models in Chapter 6.

Frequency Distributions and Bar Charts for Categorical Data

An appropriate graphical or tabular display of data can be an effective way to summarize and communicate information. When a data set is categorical, a common way to present the data is in the form of a table, called a **frequency distribution**.

● **Data set available online**

Copyright 2025 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s). Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

Definitions

Frequency distribution for categorical data: A table that displays the possible categories along with the associated frequencies and/or relative frequencies.

Frequency: The **frequency** for a particular category is the number of times the category appears in the data set.

Relative frequency: The **relative frequency** for a particular category is calculated as

$$\text{relative frequency} = \frac{\text{frequency}}{\text{number of observations in the data set}}$$

The relative frequency for a particular category is the proportion of the observations that belong to that category.

Relative frequency distribution: A frequency distribution that includes relative frequencies.

Example 1.7 Motorcycle Helmets—Can You See Those Ears?

Understand the context ➤

The U.S. Department of Transportation establishes standards for motorcycle helmets. To ensure a certain degree of safety, helmets should reach the bottom of the motorcyclist's ears. The report **"Motorcycle Helmet Use in 2019—Overall Results" (National Highway Traffic Safety Administration, June 2020)** summarized data collected in 2019 by observing 828 motorcyclists nationwide at selected roadway locations. Each time a motorcyclist passed by, the observer noted whether the rider was wearing no helmet, a noncompliant helmet, or a compliant helmet. The following coding was used:

N = no helmet
NH = noncompliant helmet
CH = compliant helmet

Consider the data ➤

A few of the observations were

CH N CH NH N CH CH CH N N

There were also 818 additional observations, which are not reproduced here. In total, there were 104 riders who wore no helmet, 138 who wore a noncompliant helmet, and 586 who wore a compliant helmet.

The corresponding frequency distribution is given in Table 1.1.

Do the work ➤

Table 1.1 Frequency Distribution for Helmet Use

Helmet Use Category	Frequency	Relative Frequency
No helmet	104	0.126 ← 104/828
Noncompliant helmet	138	0.167 ← 138/828
Compliant helmet	586	0.708 ← 586/828
	828 ← Total number of observations	1.001 ← Should total 1, but in some cases may be slightly off due to rounding

Interpret the results ➤

From the frequency distribution, you can see that about 13% of the riders were not wearing a helmet, but most of those who wore a helmet were wearing one that met the Department of Transportation safety standard.

A frequency distribution summarizes a data set in a table. It is also common to display categorical data graphically. A bar chart is one of the most widely used types of graphical displays for categorical data.

Bar Charts

A **bar chart** is a graph of a frequency distribution for categorical data. Each category in the frequency distribution is represented by a bar or rectangle, and the graph is constructed in such a way that the *area* of each bar is proportional to the corresponding frequency or relative frequency.

Bar Charts

When to Use Categorical data.

How to Construct

1. Draw a horizontal axis, and write the category names or labels below the line at equally spaced intervals.
2. Draw a vertical axis, and label the scale using either frequency or relative frequency.
3. Place a rectangular bar above each category label. The height is determined by the category's frequency or relative frequency, and all bars should have the same width. With the same width, both the height and the area of the bar are proportional to frequency and relative frequency.

What to Look For

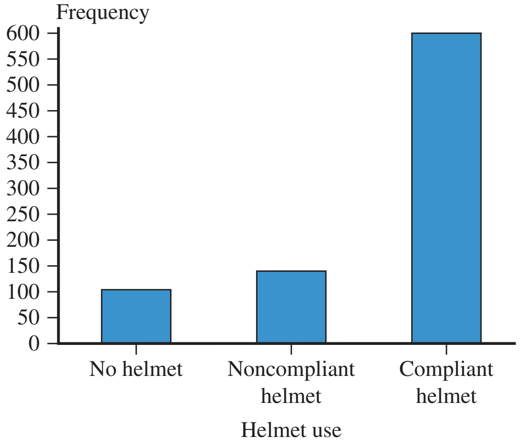
- Frequently and infrequently occurring categories.

Example 1.8 Revisiting Motorcycle Helmets

Understand the context ➤

Example 1.7 used data on helmet use from a sample of 828 motorcyclists to construct a frequency distribution (Table 1.1). Figure 1.5 shows the bar chart corresponding to this frequency distribution.

Figure 1.5
Bar chart of helmet use.



Interpret the results ➤

The bar chart provides a visual representation of the information in the frequency distribution. From the bar chart, it is easy to see that the compliant helmet use category occurred most often in the data set. The bar for compliant helmets is about four times as tall (and therefore has four times the area) as the bar for no helmet because approximately four times as many motorcyclists wore compliant helmets than wore no helmet.

Dotplots for Numerical Data

A dotplot is a simple way to display numerical data when the data set is reasonably small. Each observation is represented by a dot above the location corresponding to its value on a horizontal measurement scale (a number line). When a value occurs more than once, there is a dot for each occurrence and these dots are stacked vertically.

Dotplots

When to Use Small numerical data sets.

How to Construct

1. Draw a horizontal line and mark it with an appropriate measurement scale.
2. Locate each value in the data set along the measurement scale, and represent it by a dot. If there are two or more observations with the same value, stack the dots vertically.

What to Look for

Dotplots convey information about:

- A representative or typical value in the data set.
- The extent to which the data values vary.
- The shape of the distribution of values along the number line.
- The presence of unusual values in the data set.

Example 1.9 Making It to Graduation . . .

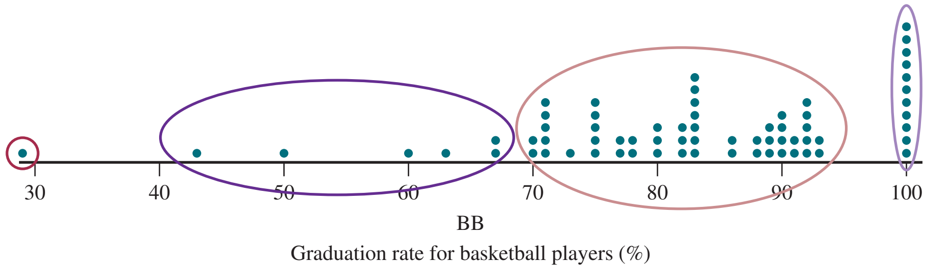
Understand the context ➤

● The article **"Keeping Score When It Counts: Academic Progress/Graduation Rate Study of the 2021 NCAA Division I Men's Basketball Tournament Teams" (The Institute for Diversity and Ethics in Sport, University of Central Florida, 2021)** compared graduation rates of basketball players to those of all student athletes for the universities and colleges that sent teams to the 2021 Division I playoffs. The graduation rates in the accompanying table are the percentage of athletes who started college in 2010, 2011, 2012, and 2013 who graduated within 6 years.

Minitab, a computer software package for statistical analysis, was used to construct a dotplot of the graduation rates for basketball players (see Figure 1.6). From this dotplot, you can see that basketball graduation rates varied a great deal from school to school, ranging from a low of 29% to a high of 100%.

You can also see that the graduation rates seem to cluster in several groups, denoted by the colored ovals that have been added to the dotplot. There are 11 schools with graduation rates of 100% (excellent!). The majority of schools are in the large cluster with graduation rates from about 70% to about 93%. And then there is a group of 6 schools with low graduation rates for basketball players and one school with an unusually low graduation rate of 29% (University of Oregon).

Figure 1.6
Minitab dotplot of graduation rates for basketball players.



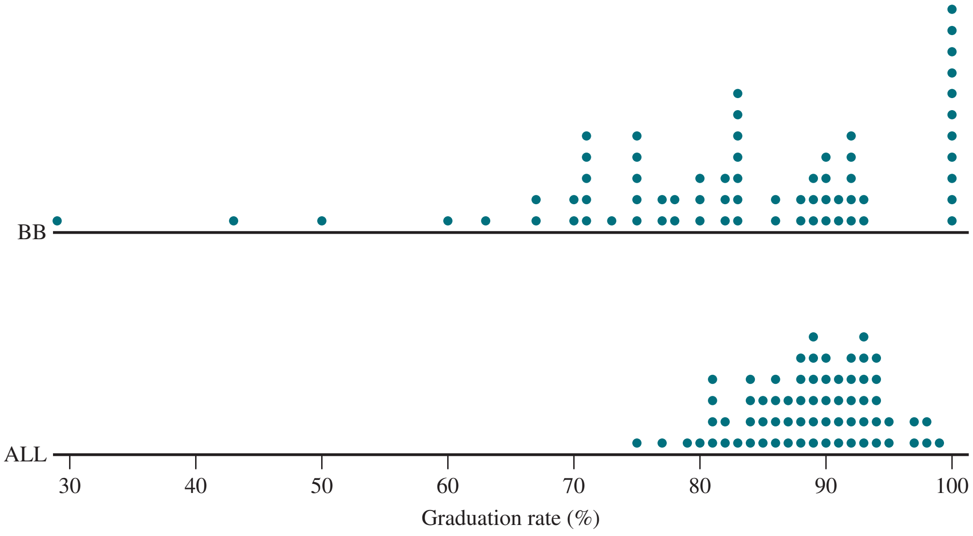
● **Data set available online**

School	BB	ALL	Difference (ALL − BB)	School	BB	ALL	Difference (ALL − BB)
1	50	84	34	35	78	80	2
2	86	85	−1	36	70	88	18
3	86	94	8	37	89	90	1
4	75	82	7	38	100	93	−7
5	100	93	−7	39	77	85	8
6	92	92	0	40	43	90	47
7	92	98	6	41	60	87	27
8	93	95	2	42	83	93	10
9	83	93	10	43	90	92	2
10	92	91	−1	44	83	89	6
11	71	77	6	45	77	94	17
12	88	81	−7	46	80	81	1
13	83	97	14	47	80	94	14
14	67	89	22	48	82	89	7
15	100	98	−2	49	83	85	2
16	71	84	13	50	82	86	4
17	100	92	−8	51	100	94	−6
18	91	88	−3	52	71	89	18
19	88	89	1	53	91	88	−3
20	90	99	9	54	75	86	11
21	100	92	−8	55	90	86	−4
22	83	79	−4	56	29	87	58
23	75	92	17	57	75	91	16
24	71	75	4	58	80	90	10
25	92	90	−2	59	78	89	11
26	71	82	11	60	100	95	−5
27	73	84	11	61	83	91	8
28	70	81	11	62	89	93	4
29	67	88	21	63	100	97	−3
30	89	90	1	64	93	81	−12
31	90	86	−4	65	75	91	16
32	100	93	−7	66	100	84	−16
33	82	94	12	67	63	87	24
34	100	83	−17	68	92	88	−4

Figure 1.7 shows two dotplots of graduation rates—one for basketball players and one for all student athletes. There are some striking differences that are easy to see when the data is displayed in this way. The graduation rates for all student athletes tend to be higher and to vary less from school to school than the graduation rates for only basketball players.

Figure 1.7

Minitab dotplot of graduation rates for basketball players (BB) and for all athletes (AA).



Copyright 2025 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s). Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

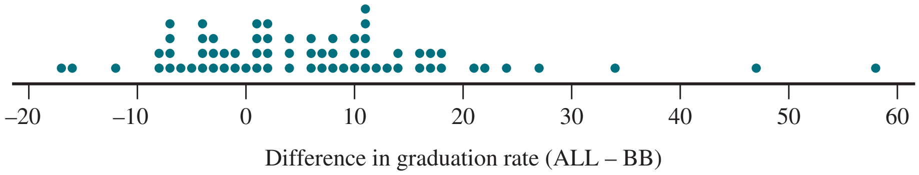
The dotplots in Figure 1.7 are informative, but you can do even better. The data given here are an example of *paired data*. Each basketball graduation rate is paired with a graduation rate for all student athletes from the same school. When data are paired in this way, it is more informative to look at differences—in this case, the difference between the graduation rate for all student athletes and for basketball players for each school. These differences (ALL – BB) are also shown in the data table.

Interpret the results ➤

Figure 1.8 gives a dotplot of the differences. Notice that one difference is equal to 0. This corresponds to a school for which the basketball graduation rate is equal to the graduation rate of all student athletes. There are 22 schools for which the difference is negative. Negative differences correspond to schools that have a higher graduation rate for basketball players than for all student athletes.

The most interesting feature of the difference dotplot is the variability in the positive differences. Positive differences correspond to schools that have a lower graduation rate for basketball players. The positive differences range from 1% all the way up to 58%. There were three schools that had a graduation rate for all athletes that was 30 percentage points or more greater than the graduation rate for basketball players. (In case you were wondering, these schools were Abilene Christian University with a difference of 24%, UCLA with a difference of 47%, and University of Oregon with a difference of 58%.)

Figure 1.8
Dotplot of graduation rate differences (all athletes – basketball players).



Exercises 1.12–1.32

● **Data set available online**

1.12 Classify each of the following variables as either categorical or numerical. For those that are numerical, determine whether they are discrete or continuous.

- **a.** Number of students in a class of 35 who turn in a term paper before the due date
- **b.** Assigned sex of the next baby born at a particular hospital
- **c.** Amount of fluid (in ounces) dispensed by a machine used to fill bottles with soda pop
- **d.** Thickness of the gelatin coating of a vitamin E capsule
- **e.** Birth order classification (only child, firstborn, middle child, lastborn) of a math major

1.13 Classify each of the following variables as either categorical or numerical. For those that are numerical, determine whether they are discrete or continuous.

- **a.** Brand of computer purchased by a customer
- **b.** State of birth for someone born in the United States
- **c.** Price of a textbook
- **d.** Concentration of a contaminant (micrograms per cubic centimeter) in a water sample
- **e.** Zip code (Think carefully about this one.)
- **f.** Actual weight of coffee in a can labeled as containing 1 pound

1.14 For the following numerical variables, state whether each is discrete or continuous.

- **a.** The number of customers who pay with cash at a grocery store on a given day
- **b.** The amount by which a 1-pound package of ground beef decreases in weight (because of moisture loss) before purchase
- **c.** The number of kernels in a bag of microwave popcorn that have not popped after 3 minutes of cooking
- **d.** The number of students in a class of 35 who have purchased a used copy of the textbook

1.15 For the following numerical variables, state whether each is discrete or continuous.

Copyright 2025 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s). Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

a. The length of a 1-year-old rattlesnake
b. The altitude of a location in California selected by throwing a dart at a map of the state
c. The distance from the left edge at which a 12-inch plastic ruler snaps when bent sufficiently to break
d. The price per gallon paid by the next customer to buy gas at a particular station

1.16 For each of the following situations, give a set of possible data values that might arise from making the observations described.
a. The manufacturer for each of the next 10 cars to pass through a given intersection is noted.
b. The grade point average for each of the 15 seniors in a statistics class is determined.
c. The number of gas pumps in use at each of 20 gas stations at a particular time is determined.
d. The actual net weight of each of 12 bags of fertilizer having a labeled weight of 50 pounds is determined.
e. Fifteen different radio stations are monitored during a 1-hour period, and the amount of time devoted to commercials is determined for each.

1.17 In a survey of 100 people who had recently purchased motorcycles, data on the following variables were recorded:
Color of motorcycle purchased
Brand of motorcycle purchased
Number of previous motorcycles owned by purchaser
Telephone area code of purchaser
Weight of motorcycle as equipped at purchase

a. Which of these variables are categorical?
b. Which of these variables are discrete numerical?
c. Which type of graphical display would be an appropriate choice for summarizing the color data, a bar chart or a dotplot?
d. Which type of graphical display would be an appropriate choice for summarizing the weight data, a bar chart or a dotplot?

1.18 Many people now use emojis like the ones shown here to communicate in text messages and email.

In a survey conducted by CustomerThermometer in 2019 **("Emoji Use in Business: A Research Report")**, people in a representative sample of 1000 Americans were asked what emoji they would use to provide feedback to a company to let them know that they thought the company was doing well. The responses are summarized in the accompanying table.

Emoji	**Relative Frequency**
Thumbs-up	0.49
Star	0.13
Smiley face	0.11
OK hand	0.05
Heart	0.04
Other	0.18

a. Use the given information to construct a bar chart.
b. There were quite a few other emojis that were lumped together in the "Other" category, but none of them occurred more often than the ones listed in the table. Given this information, how would you advise a company that was designing a method for soliciting customer feedback by clicking on an emoji symbol in an email?

1.19 ● The article **"Concussion Incidence and Trends in 20 High School Sports" (*Pediatrics*, 2019, pediatrics.aappublications.org)** reported the accompanying concussion rates for different high school sports. The given data are concussion rates per 10,000 athletes participating in high school events between 2013 and 2018.

Sport	**Concussion Rate (Concussions per 10,000athletes)**
Football	35.8
Ice Hockey	19.5
Lacrosse (Boys)	13.0
Lacrosse (Girls)	10.0
Wrestling	9.9
Basketball (Girls)	12.1
Basketball (Boys)	4.1
Soccer (Girls)	21.8
Soccer (Boys)	9.3
Field Hockey	6.5
Volleyball	5.3
Softball	3.9
Baseball	2.1

a. Construct a dotplot for the concussion rate data.
b. In addition to the three girls' sports indicated in the table (lacrosse, basketball, and soccer), the article also reported concussion rates for field hockey, volleyball, and softball, which are girls' sports. Locate the points on the dotplot that correspond to concussion rates for girls' sports and highlight them in a different color. Based on the dotplot, would you say that the concussion rates tend to be lower for girls' sports?

1.20 ● Box Office Mojo **(www.boxofficemojo.com)** tracks movie ticket sales. Ticket sales (in millions of dollars) for each of the top 20 movies in 2019 and 2020 are shown in the accompanying tables.

Movie	2019 U.S. Ticket Sales (millions of dollars)
Avengers: Endgame	858.3
The Lion King	543.6
Toy Story 4	434.0
Frozen II	430.1
Captain Marvel	426.8
Star Wars: Episode IX—The Rise of Skywalker	390.7
Spider-Man: Far from Home	390.5
Aladdin	355.6
Joker	333.8
It Chapter Two	211.6
Jumanji: The Next Level	192.1
Us	175.1
Fast & Furious Presents: Hobbs & Shaw	174.0
John Wick: Chapter 3—Parabellum	171.0
How to Train Your Dragon: The Hidden World	160.8
The Secret Life of Pets 2	157.9
Pokémon Detective Pikachu	144.1
Once Upon a Time… In Hollywood	141.1
Shazam!	140.4
Aquaman	136.0

Movie	2020 U.S. Ticket Sales (millions of dollars)
Bad Boys for Life	204.4
1917	157.9
Sonic the Hedgehog	146.1
Jumanji: The Next Level	124.7
Star Wars: Episode IX—The Rise of Skywalker	124.5
Birds of Prey	84.2
Dolittle	77.0
Little Women	70.5
The Invisible Man	64.9
The Call of the Wild	62.3
Onward	61.6
Tenet	57.2
Knives Out	49.7
Frozen II	47.2
Spies in Disguise	37.4
The Gentlemen	36.5
Just Mercy	35.7
The Croods: A New Age	32.3
Parasite	30.7
Fantasy Island	26.4

a. Construct a dotplot of the 2019 ticket sales data. Comment on any interesting features of the dotplot. (Hint: See “What to Look For” in the Dotplots box on page 14.)

b. Construct a dotplot of the 2020 ticket sales data. Comment on any interesting features of the dotplot.

c. In what ways are the distributions of the 2019 and 2020 ticket sales observations similar? In what ways are they different?

1.21 The report **“With Their Whole Lives Ahead of Them” (publicagenda.org/files/theirwholelivesaheadofthem.pdf)** includes data from a survey of 200 students who started college but did not complete a degree. Each of these students was asked, “How much have you thought about going back to school?” The accompanying frequency distribution summarizes the responses to this question.

Response	Frequency
A lot of thought	130
Some thought	48
No thought	18
Don’t know	4

a. Summarize the response data using a bar chart.

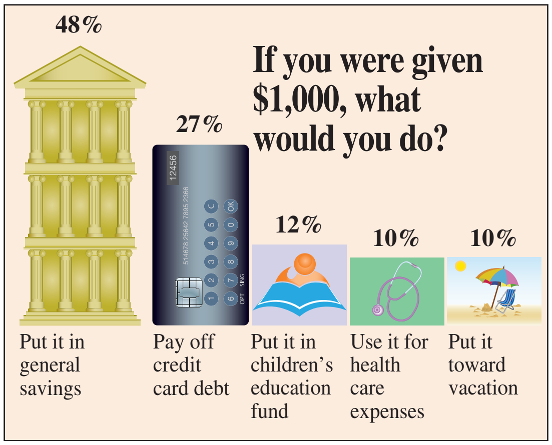
b. Write a few sentences commenting on the distribution of the responses.

1.22 The following display is a graph similar to one that appeared in **USA TODAY (June 29, 2009)**. This graph is meant to be a bar graph of responses to the question shown in the graph.

a. Is response to the question a categorical or numerical variable?

b. Explain why a bar chart rather than a dotplot was used to display the response data.

c. There must have been an error made in constructing this graph. How can you tell that the graph is not a correct representation of the response data?



Source: Bank of America Consumer Purchasing and Savings Habits survey of 1,000 adults 18 and older. Margin of error ±4 percentage points.

Copyright 2025 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s). Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

1.23 ● The following data on the number of people in prison (rounded to the nearest hundred) in 2019 for the 13 Western states are from the U.S. Department of Justice **("Correctional Populations in the United States, 2019," Bureau of Justice Statistics, 2021)**. Also included in the table is information on the state population size (Census Bureau estimates for 2019, rounded to the nearest thousand).

State	State Population	Number of People in Prison
Alaska	732,000	4,500
Arizona	7,279,000	56,000
California	39,512,000	196,100
Colorado	5,759,000	32,500
Hawaii	1,416,000	5,300
Idaho	1,787,000	12,700
Montana	1,069,000	6,900
Nevada	3,080,000	20,000
New Mexico	2,097,000	13,300
Oregon	4,218,000	21,000
Utah	3,206,000	12,500
Washington	7,615,000	30,800
Wyoming	579,000	4,000

a. Construct a dotplot using the 13 observations on number of people in prison. Which states stand out from the rest?

b. The Western states include some with a relatively small population (like Wyoming) and others with a much larger population (like California). Because of this, it might make more sense to consider incarceration *rate* by calculating the number of people in prison *per 1000 people in the state*. For example, for Alaska, the incarceration rate would be

$$\frac{4{,}500}{732{,}000}(1{,}000) = (0.0061)(1{,}000) = 6.1$$

Calculate the incarceration rates for the other 12 states. Use these values to construct a second dotplot. Do the same states stand out as unusual in this dotplot?

c. Based on your answers from Parts (a) and (b), comment on the number of people in prison for these Western states.

1.24 Heal the Bay is an environmental organization that releases an annual beach report card based on water quality **(Heal the Bay Beach Report Card, beachreportcard.org)**. The grades for 20 beaches in three counties in Washington (Whatcom, Snohomish, and Island counties) during dry weather were:

C B A+ A+ A+ A+ A+ A A+ A+
A+ C A+ A+ A+ C A F F B

Summarize the dry weather grades by constructing a relative frequency distribution and a bar chart.

1.25 The report referenced in the previous exercise also gave wet weather grades for the same beaches:

A+ A+ A+ A+ A+ A+ A+ F F F
A+ A+ F A+ A+ F A+ A+ F A+

a. Construct a bar chart for the wet weather grades.

b. Do the bar charts from Part (a) and from the previous exercise support the statement that beach water quality tends to be better in dry weather conditions? Explain.

1.26 ● The report **"Seat Belt Use in 2019" (National Highway Traffic Safety Administration)** included the estimated percentages of drivers who wear seat belts for the 50 states and the District of Columbia. In the accompanying data table, each state was also classified into one of three geographical regions—West (W), Middle states (M), or East (E).

SeatBelt%	Region	State	SeatBelt%	Region	State
92	M	AL	88	W	CO
94	W	AK	94	E	CT
91	W	AZ	93	E	DE
82	M	AR	95	E	DC
96	W	CA	90	E	FL
96	E	GA	90	E	NJ
97	W	HI	92	W	NM
86	W	ID	94	E	NY
94	M	IL	88	E	NC
95	M	IN	86	E	OH
95	M	IA	85	M	OK
85	M	KS	96	W	OR
90	M	KY	89	E	PA
88	M	LA	88	E	RI
89	E	ME	90	E	SC
90	E	MD	75	M	SD
82	E	MA	91	M	TN
94	M	MI	91	M	TX
93	M	MN	90	W	UT
81	M	MS	85	E	VA
88	M	MO	89	E	VT
89	W	MT	93	W	WA
80	M	NE	90	E	WV
94	W	NV	90	M	WI
84	M	ND	78	W	WY
71	E	NH			

a. Display the data in comparative dotplots that make it possible to compare seat belt use for the three geographical regions.

Copyright 2025 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s). Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

b. Does the graphical display in Part (a) reveal any striking differences in seat belt use for the three geographical regions or are the distributions of seat belt use similar for the three regions?

1.27 ● Example 1.5 gave the accompanying data on violent crime on public college campuses in Florida during 2016 **(fbi.gov)**:

University/College	Student Enrollment	Number of Violent Crimes Reported in 2016
Florida A&M University	9,928	6
Florida Atlantic University	30,380	6
Florida Gulf Coast University	14,833	3
Florida International University	49,782	15
Florida South Western State College	15,709	0
Florida State University, Tallahassee	40,830	20
New College of Florida	861	1
Pensacola State College	9,840	0
Santa Fe College	14,767	0
Tallahassee Community College	12,445	2
University of Central Florida	62,953	14
University of Florida	50,645	10
University of North Florida	15,675	3
University of South Florida, St. Petersburg	4,739	0
University of South Florida, Tampa	42,067	8
University of West Florida	12,763	10

a. Construct a dotplot using the 16 observations on number of violent crimes reported. Which schools stand out from the rest?

b. One of the Florida schools has only 861 students and a few of the schools are quite a bit larger than the rest. Because of this, it might make more sense to consider crime rate by calculating the number of violent crimes reported per 1000 students. For example, for Florida A&M University the violent crime rate would be

$$\frac{6}{9928}(1000) = (0.001)(1000) = 1.0$$

Calculate the violent crime rates for the other 15 schools and then use those values to construct a dotplot. Do the same schools stand out as unusual in this dotplot?

c. Based on your answers from Parts (a) and (b), write a couple of sentences commenting on violent crimes reported at Florida universities and colleges in 2016.

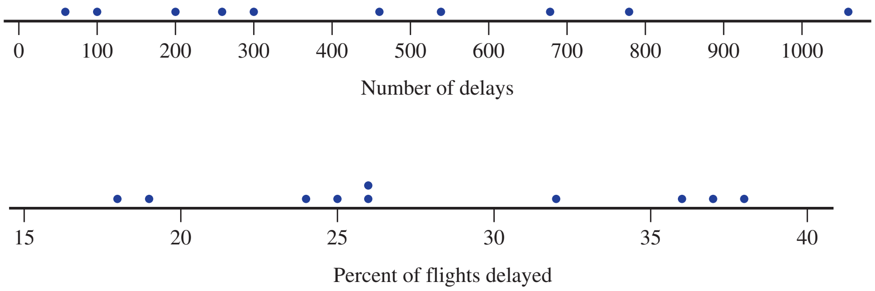
1.28 ● FlightAware **(flightaware.com)** publishes data on airline delays and cancelations. Data on delayed flights on December 30, 2022, for 10 U.S. airlines are given in the accompanying table.

Airline	Number of Delays	Percent of Flights Delayed
Delta	541	18
American	783	24
Alaska	303	38
Hawaiian	93	36
Spirit	262	32
Mesa	61	19
United	674	26
Frontier	199	37
SkyWest	467	25
Southwest	1064	26

The graph at the bottom of the page shows two dotplots. One displays the data on number of delays and the other displays the data on the percentage of all flight that were delayed.

a. If you were going to rank airlines based on these flight delay data, would you use the total number of flights delayed or the percentage of flights delayed? Explain the reason for your choice.

Figure for Exercise 1.28



Copyright 2025 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s). Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

b. Write a short paragraph that could be used as part of a newspaper article on flight delays that could accompany these two dotplots.

1.29 The Gallup report **"Country Living Enjoys Renewed Appeal in U.S."** included data from a poll of a random sample of 1018 adults living in the United States in 2020. Each person in the sample was asked the following question: If you could live anywhere you wished, where would you prefer to live in a big city, a small city, a suburb of a big city, a suburb of a small city, a town, or a rural area? The responses are summarized in the given relative frequency distribution.

Response	Relative Frequency
Big city	0.11
Small city	0.16
Suburb of a big city	0.16
Suburb of a small city	0.09
Town	0.17
Rural area	0.31

a. Use this information to construct a bar chart for the response data.

b. Comment on the distribution of responses displayed in the bar chart.

1.30 Students at SUNY Oneonta participated in a survey designed to find out how they used various course resources **("Student Textbook and Course Material Survey," 2021)**. The survey was sent to 1500 students and 165 of these students completed the survey. Responses to a question asking what resource they turned to most often when they had a question about course content are summarized in the accompanying table.

Resource	Frequency
Course management system (Blackboard)	49
Google	38
Instructor	36
Classmates	15
Other	27

a. Use the given information to construct bar chart.

b. Write a few sentences commenting on resources students use to find answers to questions about course content.

1.31 The report **2013 International Bedroom Poll: Summary of Findings** describes a survey of 251 adult Americans conducted by the National Sleep Foundation **(sleepfoundation.org)**. Participants in the survey were asked how often they change the sheets on their bed and were asked to respond with one of the following categories: more than once a week, once a week, every other week, every 3 weeks, or less often than every 3 weeks. For this group, 10% responded more than once a week, 53% responded once a week, 26% responded every other week, 5% responded every 3 weeks, and 6% responded less often than every 3 weeks.

a. Use the given information to make a relative frequency distribution for the responses to the question.

b. Summarize the given information by constructing a bar chart.

1.32 In the United States, movies are rated by the Motion Picture Association of America (MPAA). The accompanying table gives the MPAA rating of the 25 top money-making movies of 2019. (data from **www.the-numbers.com/market/2019/top-grossing-movies**).

Movie (Ordered from high to low based on amount of money made)	Rating
Avengers: Endgame	PG-13
The Lion King	PG
Frozen II	PG
Toy Story 4	G
Captain Marvel	PG-13
Star Wars: The Rise of Skywalker	PG-13
Spider-Man: Far From Home	PG-13
Aladdin	PG
Joker	R
It: Chapter Two	R
Jumanji: The Next Level	PG-13
Us	R
Fast and Furious Presents: Hobbs & Shaw	PG-13
John Wick: Chapter 3—Parabellum	R
How to Train Your Dragon: The Hidden World	PG
The Secret Life of Pets 2	PG
Pokémon: Detective Pikachu	PG
Once Upon a Time…In Hollywood	R
Shazam!	PG-13
Aquaman	PG-13
Knives Out	PG-13
Dumbo	PG
Maleficent: Mistress of Evil	PG
Glass	PG-13
Godzilla: King of the Monsters	PG-13

Use the given information to construct a bar chart of the ratings for the top 25 movies of 2019. Write a few sentences describing the ratings distribution.

Chapter Activities

Activity 1.1 Head Sizes: Understanding Variability

The purpose of this activity is to develop an initial understanding of variability.

Materials needed: Each team will need a measuring tape. For this activity, you will work in teams of 6 to 10 people.

1. Designate a team leader for your team by choosing the person on your team who celebrated their last birthday most recently.
2. The team leader should measure and record the head size (measured as the circumference at the widest part of the forehead) of each of the other members of their team.
3. Record the head sizes for the individuals on your team as measured by the team leader.
4. Next, each individual on the team should measure the head size of the team leader. Do not share your measurement with the other team members until all team members have measured the team leader's head size.
5. After all team members have measured the team leader's head, record the different team leader head size measurements obtained by the individuals on your team.
6. Using the data from Step 3, construct a dotplot of the team leader's measurements of team head sizes. Then, using the same scale, construct a separate dotplot of the different measurements of the team leader's head size (from Step 5).

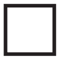
Now use the available information to answer the following questions:

7. Do you think the team leader's head size changed in between measurements? If not, explain why the measurements of the team leader's head size are not all the same.
8. Which data set was more variable—head size measurements of the different individuals on your team or the different measurements of the team leader's head size? Explain the basis for your choice.
9. Consider the following scheme (you don't actually have to carry this out): Suppose that a group of 10 people measured head sizes by first assigning each person in the group a number between 1 and 10. Then person 1 measured person 2's head size, person 2 measured person 3's head size, and so on, with person 10 finally measuring person 1's head size. Do you think that the resulting head size measurements would be more variable, less variable, or show about the same amount of variability as a set of 10 measurements resulting from a single individual measuring the head size of all 10 people in the group? Explain.

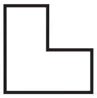
Activity 1.2 Estimating Shape Sizes

The purpose of this activity is to introduce the concept of accuracy of an estimate.

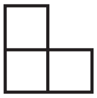
1. Construct an activity sheet that consists of a table that has 6 columns and 10 rows. Label the columns of the table with the following six headings: (1) Shape Number, (2) Estimated Size, (3) Actual Size, (4) Difference (Estimated − Actual), (5) Absolute Difference, and (6) Squared Difference. Enter the numbers from 1 to 10 in the "Shape" column.
2. Next you will be visually estimating the sizes of the shapes in Figure 1.9. Size will be described as the number of squares of this size



that would fit in the shape. For example, the shape



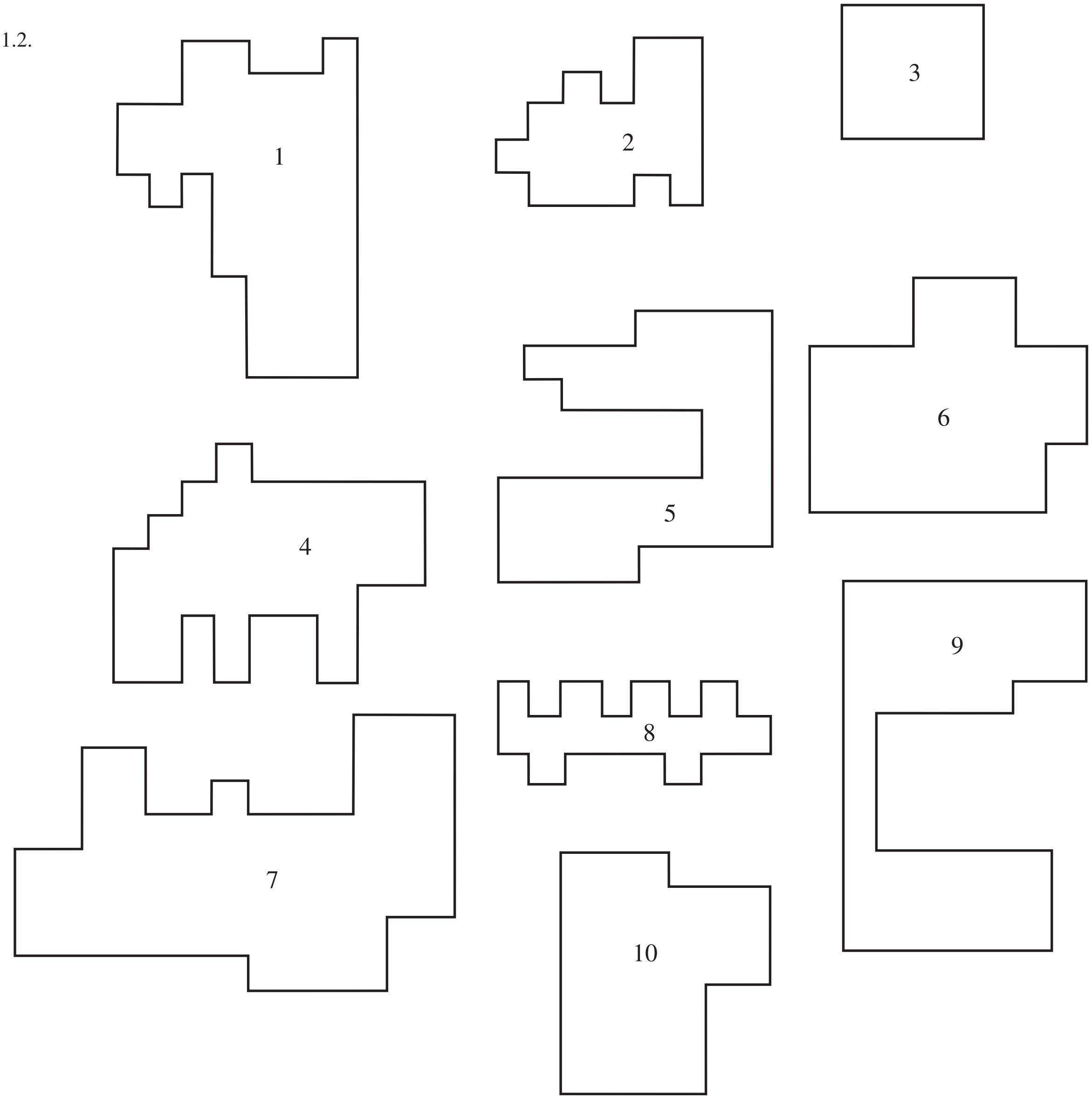
would be size 3, as illustrated by



You should now quickly *visually* estimate the sizes of the shapes in Figure 1.9. *Do not* draw on the figure—these are to be quick visual estimates. Record your estimates in the "Estimated Size" column of the activity sheet.
3. Your instructor will provide the actual sizes for the 10 shapes, which should be entered into the "Actual Size" column of the activity sheet. Now complete the "Difference" column by subtracting the actual value from your estimate for each of the 10 shapes.
4. What would cause a difference to be negative? What would cause a difference to be positive?
5. Would the sum of the differences tell you if the estimates and actual values were in close agreement?

Copyright 2025 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s). Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

Figure 1.9
Shapes for Activity 1.2.



Does a sum of 0 for the differences indicate that all the estimates were equal to the actual value? Explain.

6. Compare your estimates with those of another person in the class by comparing the sum of the absolute values of the differences between estimates and corresponding actual values. Who was better at estimating shape sizes? How can you tell?
7. Use the last column of the activity sheet to record the squared differences (for example, if the difference for shape 1 was -3, the squared difference would be $(-3)^2 = 9$). Explain why the sum of the squared differences can also be used to assess how accurate your shape estimates were.
8. For this step, work with three or four other students from your class. For each of the 10 shapes, form a new size estimate by calculating the average of the size estimates for that shape made by the individuals in your group. Is this new set of estimates more accurate than your own individual estimates were? How can you tell?
9. Does your answer from Step 8 surprise you? Explain why or why not.

Copyright 2025 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s). Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

Activity 1.3 A Meaningful Paragraph

Write a meaningful paragraph that includes the following six terms: **sample, population, descriptive statistics, bar chart, numerical variable,** and **dotplot**.

A "meaningful paragraph" is a coherent piece of writing in an appropriate context that uses all of the listed words. The paragraph should show that you understand the meanings of the terms and their relationships to one another. A sequence of sentences that just defines the terms is *not* a meaningful paragraph. When choosing a context, think carefully about the terms you need to use. Choosing a good context will make writing a meaningful paragraph easier.

Summary Key Concepts and Formulas

Term or Formula	Comment
Population	The entire collection of individuals or measurements about which information is desired.
Sample	A part of the population selected for study.
Descriptive statistics	Numerical, graphical, and tabular methods for organizing and summarizing data.
Inferential statistics	Methods for generalizing from a sample to a population.
Categorical data	Individual observations are categorical responses (nonnumerical).
Numerical data	Individual observations are numerical (quantitative) in nature.
Discrete numerical data	Possible values are isolated points along the number line.
Continuous numerical data	Possible values form an entire interval along the number line.
Univariate, bivariate, and multivariate data	Each observation consists of one (univariate), two (bivariate), or two or more (multivariate) responses or values.
Frequency distribution for categorical data	A table that displays frequencies, and sometimes relative frequencies, for each of the possible values of a categorical variable.
Bar chart	A graph of a frequency distribution for a categorical data set. Each category is represented by a bar, and the area of the bar is proportional to the corresponding frequency or relative frequency.
Dotplot	A graph of numerical data in which each observation is represented by a dot on or above a horizontal measurement scale.

Chapter Review Exercises 1.33–1.38

● Data set available onlinee

1.33 ● The report "Traffic Safety Facts" (National Highway Traffic Safety Administration, June 2015) gave data on traffic fatality rate (traffic deaths per 100,000 population) for the 50 U.S. states and the District of Columbia. Some, but not all, states enforce seat belt laws. Below are the fatality rate data divided into two groups—states that enforce seat belt laws and those that do not.

States with Seat Belt Law Enforcement

17 9 18 8 7 13 3 15 14 7 8 12 10
12 17 16 12 9 10 7 23 6 14 6 14 16
11 4 20 15 13 8 15 10

States without Seat Belt Law Enforcement

13 10 13 5 14 22 13 11 9 17 10 9 15
9 9 9 25

a. Construct a dotplot of the fatality rates for states that have seat belt law enforcement. Write a few sentences describing any interesting features of the dotplot.

b. Construct a dotplot of the fatality rates for states that do not have seat belt law enforcement. Write a few sentences describing any interesting features of the dotplot.

c. Comment on similarities and differences in the fatality rate distributions for states with seat belt enforcement and states without seat belt enforcement.

1.34 The National Center for Education Statistics (nces,ed.gov) reported that 22% of adults were classified as being below a basic literacy level, 32% were classified as being at a basic literacy level, 46% were classified as being at a proficient level.

a. Is the variable *literacy level* categorical or numerical?

b. Construct a bar chart to display the given data on literacy level.

c. Would it be appropriate to display the given information using a dotplot? Explain why or why not.

Copyright 2025 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s). Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

1.35 ● The Computer Assisted Assessment Center at the University of Luton published a report titled **"Technical Review of Plagiarism Detection Software."** The authors of this report asked faculty at academic institutions about the extent to which they agreed with the statement "Plagiarism is a significant problem in academic institutions." The responses are summarized in the accompanying table. Construct a bar chart for these data.

Response	Frequency
Strongly disagree	5
Disagree	48
Not sure	90
Agree	140
Strongly agree	39

1.36 In 2018, researchers in Florida surveyed a large sample of students at Florida colleges and universities to learn about how students use textbooks and textbook costs **(*"2018 Student Textbook and Course Materials Survey,"* March 8, 2019).** One question on the survey asked students if they were willing to rent textbooks. The accompanying frequency table summarizes the student responses.

Response	Frequency
Yes, either printed or digital	12,635
Yes, only if printed	5,497
Yes, only if digital	536
No	1,039
Maybe	1,723

Construct a bar chart to summarize the data distribution. Write a few sentences commenting on the students' willingness to rent textbooks.

1.37 ● Each year, ***U.S. News and World Report*** publishes a ranking of U.S. business schools. The following data give the acceptance rates (percentage of applicants admitted) for the best 25 programs in a recent survey:

16.3 12.0 25.1 20.3 31.9 20.7 30.1 19.5 36.2
46.9 25.8 36.7 33.8 24.2 21.5 35.1 37.6 23.9
17.0 38.4 31.2 43.8 28.9 31.4 48.9

a. Construct a dotplot.

b. Comment on the interesting features of the plot.

1.38 The report **"The American Freshman: National Norms Fall 2019" (Higher Education Research Institute, 2019)** summarizes data collected from more than 95,000 college freshmen across the United States. One question on the survey asked students how they identified politically, with choices of far left, liberal, middle-of-the-road, conservative, and far right. The data are summarized in the accompanying table.

Use the given data to construct a bar chart. Write a brief description of how the surveyed students identified politically that is based on the bar chart.

Response	Relative Frequency
Far left	0.045
Liberal	0.322
Middle-of-the-road	0.436
Conservative	0.178
Far right	0.019

Technology Notes

Bar Charts

JMP

1. Enter the raw data into a column (**Note:** To open a new data table, click **File** then select **New** then **Data Table**)
2. Click **Graph**, **Legacy** and then select **Chart**
3. Click and drag the column name containing your stored data from the box under **Select Columns** to the box next to **Categories, X, Levels**
4. Click **OK**

Minitab

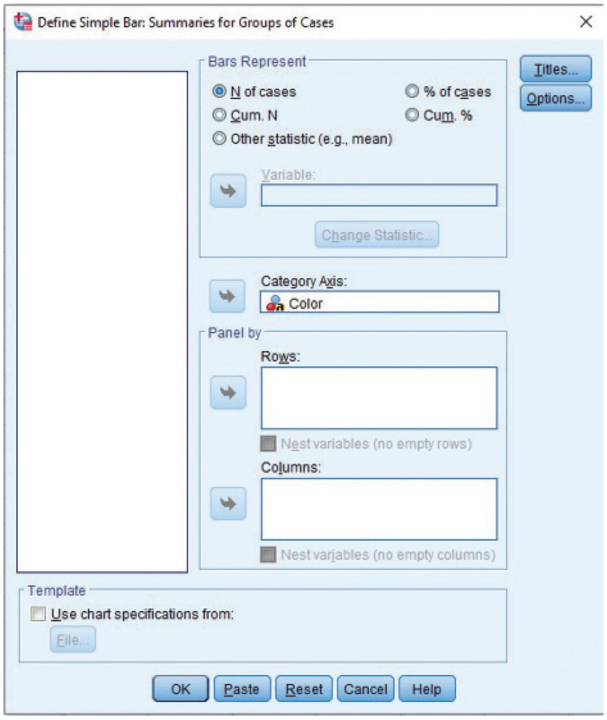
1. Enter the raw data into C1
2. Select **Graph** and choose **Bar Chart…**
3. Highlight Simple
4. Click **OK**
5. Double click C1 to add it to the **Categorical Variables** box
6. Click **OK**

Note: You may add or format titles, axis titles, legends, etc., by clicking on the **Labels…** button prior to performing step 6 above.

SPSS

1. Enter the raw data into a column
2. Select **Graph** and Legacy Dialogs and choose **Bar…**
3. Click on Simple
4. Click **Define**

Copyright 2025 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s). Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.


Define Simple Bar: Summaries for Groups of Cases
Bars Represent
N of cases
% of cases
Cum. N
Cum. %
Other statistic (e.g., mean)
Variable:
Change Statistic...
Titles...
Options...
Panel by
Rows:
Nest variables (no empty rows)
Columns:
Nest variables (no empty columns)
Template
Use chart specifications from:
File...
OK
Paste
Reset
Cancel
Help

● N of cases

Category Axis: Color

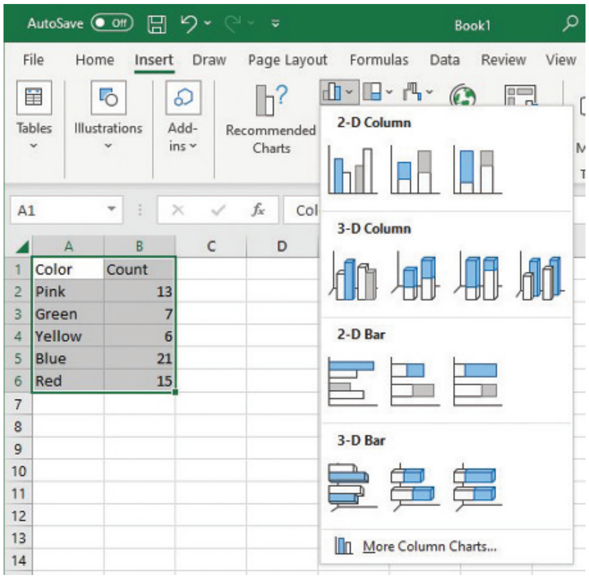
Source: SPSS

5. Click the arrow next to **Category Axis**: to import the variable to be graphed
6. Click **OK**

Note: By default, SPSS displays the count of each category in a bar chart. To display the percentage for each category, follow the above instructions to Step 5. Then, in the Element Properties dialog box, select Percentage from the drop-down box under Statistic. Click **Apply** then click **OK** on the Chart Builder dialog box.

Excel

1. Enter the category names into column A (you may input the title for the variable in cell A1)
2. Enter the count or percent for each category into column B (you may put the title "Count" or "Percentage" in cell B1)
3. Select all data (including column titles if used)
4. Click on the **Insert** Ribbon


AutoSave Off
Book1
File
Home
Insert
Draw
Page Layout
Formulas
Data
Review
View
Tables
Illustrations
Add-ins
Recommended Charts
2-D Column
3-D Column
2-D Bar
3-D Bar
More Column Charts...
A1

	A	B	C	D
1	Color	Count		
2	Pink	13		
3	Green	7		
4	Yellow	6		
5	Blue	21		
6	Red	15		

Source: Microsoft Corporation

5. Choose the **Insert Column or Bar Chart** icon in the **Charts** section of the ribbon then select the first chart under 2-D column (Clustered Column)
6. The chart will appear on the same worksheet as your data

Note: You may add or format titles, axis titles, legends, and so on, by right-clicking on the appropriate piece of the chart.

Note: Using the Bar Option on the Insert Ribbon will produce a horizontal bar chart.

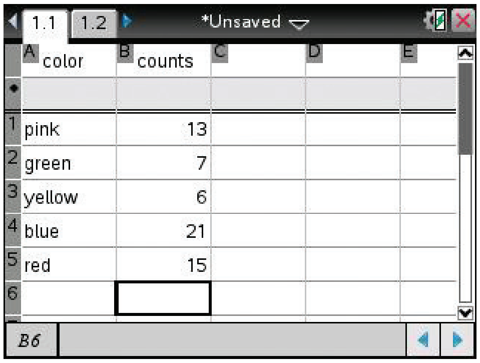
TI-83/84

The TI-83/84 does not have the functionality to produce bar charts.

TI-Nspire CX

1. Enter the category names into a list (to access data lists select the spreadsheet option and press **enter**)

Note: Be sure to title the list by selecting the top row of the column and typing a title.


1.1
1.2
*Unsaved
B6

	A color	B counts	C	D	E
1	pink	13			
2	green	7			
3	yellow	6			
4	blue	21			
5	red	15			
6					

Source: TI-Nspire CX

2. Enter the counts for each category into a second list (be sure to title this list as well!)
3. Press the **menu** key and navigate to **3:Data** and then **8:Summary Plot** and select this option
4. For **X List** select the list containing your category names from the drop-down menu
5. For **Summary List** select the list containing the counts for each category from the drop-down menu
6. For **Display On** select either **Split Page** or **New Page** from the drop-down menu
7. Press **OK**

Dotplots

JMP

1. Input the raw data into a column
2. Click **Graph** and select **Chart**
3. Click and drag the column name containing the data from the box under **Select columns** to the box next to **Categories, X, Levels**
4. Select **Point Chart** from the second drop-down box in the **Options** section

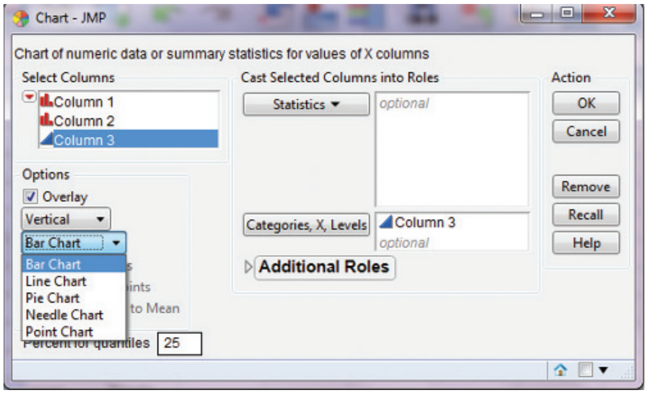


Chart - JMP
Chart of numeric data or summary statistics for values of X columns
Select Columns
Column 1
Column 2
Column 3
Cast Selected Columns into Roles
Statistics
optional
Action
OK
Cancel
Remove
Recall
Help
Options
Vertical
Bar Chart
Line Chart
Pie Chart
Needle Chart
Point Chart
ints
to Mean
Additional Roles

☑ Overlay

Categories, X, Levels: Column 3

Percent for quantiles: 25

Source: JMP

5. Click **OK**

Minitab

1. Input the raw data into C1
2. Select **Graph** and choose **Dotplot**
3. Highlight Simple under One Y
4. Click **OK**
5. Double click C1 to add it to the **Graph Variables** box
6. Click **OK**

Note: You may add or format titles, axis titles, legends, and so on, by clicking on the **Labels…** button prior to performing Step 6 above.

SPSS

1. Input the raw data into a column
2. Select **Graph** and choose **Chart Builder…**
3. Under **Choose from** highlight Scatter/Dot
4. Click and drag the second option in the second row (Simple Dot Plot) to the Chart preview area
5. Click and drag the variable name from the **Variables** box into the **X-Axis?** box
6. Click **OK**

Excel

Excel does not have the functionality to create dotplots.

TI-83/84

The TI-83/84 does not have the functionality to create dotplots.

TI-Nspire

1. Enter the data into a data list (to access data lists select the spreadsheet option and press **enter**)

Note: Be sure to title the list by selecting the top row of the column and typing a title.

2. Press the **menu** key then select **3:Data** then select **9:QuickGraph** and press **enter** (a dotplot will appear)

Copyright 2025 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s). Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

Chapter 2

Collecting Data Sensibly

KOTOIMAGES/Shutterstock.com

Data and conclusions based on data are everywhere—in newspapers, magazines, online resources, and professional publications. But should you believe what you read? For example, should you supplement your diet with black currant oil to stop hair loss? Will playing solitaire for 20 minutes each day help you feel less tired? If you eat proteins before carbohydrates when you eat a meal, will it lower your blood sugar? Should you donate blood twice a year to lower your risk of heart disease? These are just four recommendations out of many that appear in one issue of ***Woman's World*** **(April 4, 2016)**, a magazine with more than 1.3 million readers. In fact, if you followed all of the recommendations in that issue, you would also be loading up on prickly pear oil, hot chocolate, ginger tea, bread, bananas, sweet potatoes, bell peppers, tomatoes, and onions! The magazine suggests that these claims are based on research studies, but how reliable are the studies? Are the conclusions drawn reasonable, and do they apply to you?

A primary goal of statistical studies is to collect data that can be used to make informed decisions. It should come as no surprise that the ability to make good decisions depends on the quality of the information available.

Both the type of analysis that is appropriate and the conclusions that can be drawn depend on how the data are collected. In this chapter, two types of statistical studies and two widely used methods of data collection (sampling and experimentation) are considered.

Learning Objectives

After completing this chapter, you will be able to:

- Distinguish between an observational study and an experiment.
- Describe how the types of conclusions that can be drawn from data depend on the way that the data were collected.
- Distinguish between selection bias, measurement or response bias, and nonresponse bias.
- Explain why random assignment is an important component of a sampling plan.
- Select a simple random sample from a given population.
- Distinguish between simple random sampling, stratified random sampling, cluster sampling, systematic sampling, and convenience sampling.

- Explain why random assignment is important when collecting data in an experiment.
- Describe a procedure for randomly assigning subjects to treatments in an experiment.
- Design a completely randomized experiment.
- Explain the purpose of a control group and of blinding in an experiment.
- Design a randomized block experiment.

Section 2.1 Statistical Studies: Observation and Experimentation

In 2015, a study published in *The Lancet* described the relationship between breastfeeding in the first year of life and a person's intelligence, educational attainment, and income at 30 years of age. The results of this study were reported by a number of different news media. Some of the headlines that appeared follow:

> **"More Evidence Breastfeeding Makes Babies Smarter" (*Today*)**
> **"Breast-Feeding Boosts Chances of Success, Study in Brazil Finds" (NPR)**
> **"Benefits of Breastfeeding: Babies Grow Up to Be Smarter, High Earning Adults" (organicauthority.com)**

The study that was the basis for these headlines involved over 3000 participants, all born in Pelotas, Brazil. Information about breastfeeding was recorded in the participants' early childhood. At the age of 30, the participants were asked to take an IQ test and answer survey questions about their educational attainment and monthly income. The study found that participants who were breastfed for 12 months or more had higher IQ scores (3.76 points, on average), more years of education (0.91 years, on average), and higher monthly incomes (341 Brazilian reals, on average). The headlines shown imply that breastfeeding was the cause of these observed differences. Is that conclusion reasonable?

The answer depends in a critical way on the study design. After considering some important aspects of study design, these headlines will be revisited to consider whether they are appropriate.

Observation and Experimentation

Data collection is an important step in the data analysis process. When you set out to collect information, it is important to keep in mind the questions you hope to answer on the basis of the resulting data. Sometimes you are interested in answering questions about characteristics of a single existing population or in comparing two or more well-defined populations. To accomplish this, you could select a sample from each population under consideration and use the sample information to gain insight into characteristics of those populations.

For example, an ecologist might be interested in estimating the average shell thickness of bald eagle eggs. A social scientist studying a rural community may want to determine whether age and attitude toward abortion are related. These are examples of studies that are *observational* in nature. In these studies, you want to observe characteristics of members of an existing population or of several populations, and then use the resulting information to draw conclusions. In **observational studies**, it is important to obtain samples that are representative of the corresponding populations.

Sometimes the questions you are trying to answer deal with the effect of certain explanatory variables on some response and cannot be answered using data from an observational study. These questions are often of the form, "What happens when . . . ?" or, "What is the effect of . . . ?" For example, an educator may wonder what would happen to test scores if the required lab time for a chemistry course was increased from 3 hours to 6 hours per week. To answer such questions, an experiment might be carried out to collect relevant data. The value of some response variable (test score in the chemistry example) could be recorded under different experimental conditions (3-hour lab and 6-hour lab). In an **experiment**, one or more explanatory variables, also sometimes called factors, define the experimental conditions.

The type of conclusion that can be drawn from a statistical study depends on how the study was conducted. Both observational studies and experiments can be used to compare groups, but in an experiment the researcher controls who is in which group, whereas this is not the case in an observational study. This seemingly small difference is critical when it comes to drawing conclusions based on data from the study.

Definitions

Observational study: A study in which characteristics of a sample selected from one or more existing populations are observed. The goal of an observational study is usually to draw conclusions about the corresponding population or about differences between two or more populations. In a well-designed observational study, the sample is selected in a way that is designed to produce a sample that is representative of the population.

Experiment: A study that investigates how a response variable behaves when one or more explanatory variables, also called factors, are manipulated. The usual goal of an experiment is to determine the effect of the explanatory variables (factors) on the response variable. In a well-designed experiment, the composition of the groups that will be exposed to different experimental conditions is determined by random assignment.

A well-designed experiment can result in data that provide evidence for a cause-and-effect relationship. This is an important difference between an observational study and an experiment. In an observational study, it is not possible to draw clear cause-and-effect conclusions because you cannot rule out the possibility that the observed effect is due to some variable other than the explanatory variable being studied. Such variables are called **confounding variables**.

Definition

Confounding variable: A variable that is related to both how the experimental groups were formed and the response variable of interest.

Consider the role of confounding variables in the following two studies:

- The paper **"Living Near Noise Pollution Tied to Greater Risk of Heart Attack" (*American College of Cardiology*, March 23, 2022)** summarizes a study of the relationship between noise exposure and heart disease. Researchers compared heart attack rates (heart attacks per 100,000 people) for people who lived in places with high transportation noise exposure and people who lived in quieter areas. They concluded that the heart attack rate was 72% greater for those who lived in the high noise areas. The paper notes that living near roadways and other transportation infrastructure could also mean greater exposure to air pollution and might also be related to other health and socioeconomic factors. These represent potential confounding variables, and so it is not reasonable to draw the conclusion that noise exposure was the cause of the increased heart attack rate based on this observational study.
- Studies have shown that people over age 65 who get a flu shot are less likely to die from a flu-related illness during the following year than those who do not get a flu shot. However, recent research has shown that people over age 65 who get flu shots are also less likely to die from *any* cause during the following year than those who don't get flu shots **("Flu Vaccine and Mortality in Hypertension: A National Cohort Study," *Journal of the American Heart Association* [2022])**. This has led to the speculation that those over age 65 who get flu shots may be healthier as a group than those who do not get flu shots. If this is the case, observational studies that compare two groups—those who get flu shots and those who do not—may overestimate the effectiveness of the flu vaccine because general health differs in the two groups. General health is a possible confounding variable in such studies.

Each of the studies described above illustrates why potential confounding variables make it unreasonable to draw cause-and-effect conclusions from observational studies.

Let's return to the study on breastfeeding described at the beginning of this section. Is this an observational study or an experiment? Two groups were compared (children who were breastfed for at least a year and children who were not), but the researchers did not randomly assign children to the groups. The study is observational, and so cause-and-effect conclusions such as "breastfeeding makes babies smarter" are not justified. There is evidence that, as a group, children who were breastfed for over a year tend to have better outcomes than those who were not. However, you should not say that breastfeeding is the *cause* of these outcomes. It is possible that other variables—such as socioeconomic status, parents' IQ and education level, or early childhood education—may be related to both breastfeeding and outcomes like IQ, educational attainment, and income.

Fortunately, not everyone made the same mistake as the writers of the headlines given earlier in this section. For example, the **BBC** used the headline **"Breastfeeding Linked to Higher IQ,"** which does not imply a cause-and-effect relationship.

Drawing Conclusions from Statistical Studies

In this section, two different types of conclusions have been described. One type involves generalizing from what has been seen in a sample to some larger population, and the other involves reaching a cause-and-effect conclusion about the effect of an explanatory variable on a response. When is it reasonable to draw such conclusions? The answer depends on the way that the data were collected. Table 2.1 summarizes the types of conclusions that can be made with different study designs.

As can be seen from Table 2.1, it is important to think carefully about the objectives of a statistical study before planning how data will be collected. Both observational studies and experiments must be carefully designed if the resulting data are to be useful. The common sampling procedures used in observational studies are considered in Section 2.2. In Sections 2.3 and 2.4, you will consider experimentation and explore what constitutes good practice in the design of simple experiments.

Table 2.1 Drawing Conclusions from Statistical Studies

Study Description	Reasonable to Generalize Conclusions about Group Characteristics to the Population?	Reasonable to Draw Cause-and-Effect Conclusion?
Observational study with sample selected at random from population of interest	Yes	No
Observational study based on convenience or voluntary response sample (poorly designed sampling plan)	No	No
Experiment with groups formed by random assignment of individuals or objects to experimental conditions		
• Individuals or objects used in study are volunteers or not randomly selected from some population of interest	No	Yes
• Individuals or objects used in study are randomly selected from some population of interest	Yes	Yes
Experiment with groups not formed by random assignment to experimental conditions (poorly designed experiment)	No	No

Exercises 2.1–2.12

2.1 The article **"Walking Dogs Is Sending Older People to the ER, Study Says" (cnn.com, March 6, 2019)** describes a study that analyzed data from hospital emergency departments throughout the United States from 2004 to 2017. The researchers found over 32,000 cases of fractures among people 65 and older that were associated with walking dogs. They also found that these cases were becoming more frequent over time. Is this study an observational study or an experiment? Give a brief explanation for your choice.

Copyright 2025 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s). Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

2.2 The paper **"The Impact of Information Disclosure on Consumer Behavior" (National Bureau of Economic Research, 2018)** described a study of the ordering behavior of diners at full-service restaurants on a college campus. When parties of diners checked in to be seated, some parties received the usual set of menus, and some parties received identical menus with the addition of calorie labels. The researchers were interested in whether receiving information about the number of calories in each menu item would influence what the diners ordered. The researchers summarized their conclusions with the following statement: "We estimate that the labels resulted in a 3.0% reduction in calories ordered, with the reduction occurring in appetizers and entrees but not drinks or desserts." Is this study an observational study or an experiment? Give a brief explanation for your choice.

2.3 The authors of the paper **"What's in a Name: Exposing Gender Bias in Student Ratings of Teaching" (*Innovative Higher Education* [2015]: 291–303)** carried out a study to see whether instructor gender could have an effect on student evaluations. The course professor assigned students to one of four discussion groups taught by two assistant instructors—one who identified as a man and one who identified as a woman. However, each instructor taught one group under their own gender identity and one group under the other instructor's gender identity, as shown in the table below.

Discussion Group	Instructor's Perceived Gender	Instructor's Actual Gender
A	Woman	Woman
B	Woman	Man
C	Man	Woman
D	Man	Man

The study found that the average rating of the instructors perceived to be men was higher than the average rating of the instructors perceived to be women, regardless of actual gender. Indicate whether the study is an observational study or an experiment. Give a brief explanation for your choice.

2.4 **"Study Says Coffee May Reduce Risk of Liver Disease"** was a headline that appeared in ***USA Today* (June 23, 2021)**. This conclusion was based on data from 494,000 people in the UK Biobank, a biomedical database. The study monitored participants' liver health over an 11-year period and found that coffee drinkers were 21% less likely to develop chronic liver disease and 49% less likely to die of chronic liver disease than non-coffee drinkers.

a. Based on this study, is it reasonable to conclude that drinking coffee causes lower risk of liver disease? Explain.

b. Is it reasonable to generalize the results of this study to all American adults? Explain why or why not. (Hint: Look at Table 2.1.)

2.5 The article **"Acupuncture for Bad Backs: Even Sham Therapy Works" (*Time,* May 12, 2009)** summarized a study conducted by researchers in Seattle. In this study, 638 adults with back pain were randomly assigned to one of four groups. People in group 1 received the usual care for back pain. People in group 2 received acupuncture at a set of points tailored specifically for each individual. People in group 3 received acupuncture at a standard set of points typically used in the treatment of back pain. Those in group 4 received fake acupuncture—they were poked with a toothpick at the same set of points chosen for the people in group 3!

Two conclusions from the study were: (1) patients receiving real or fake acupuncture experienced a greater reduction in pain than those receiving usual care and (2) there was no significant difference in pain reduction for those who received acupuncture and those who received fake acupuncture toothpick pokes.

a. Is this study an observational study or an experiment? Explain.

b. Is it reasonable to conclude that receiving either real or fake acupuncture was the cause of the observed reduction in pain in those groups compared to the usual care group? What aspect of this study supports your answer? (Hint: Look at Table 2.1.)

2.6 **"Diet Rich in Omega-3s Linked to Fewer Migraines, Less Pain"** is a headline that appeared on the website Medscape **(www.medscape.com, July 6, 2021)**. Consider the following five possible study descriptions. For each of the study descriptions, answer these five questions:

Question 1: Is the described study an observational study or an experiment?

Question 2: Did the study use random selection from some population?

Question 3: Did the study use random assignment to experimental groups?

Question 4: Would it be reasonable to generalize conclusions from this study to some larger population? If so, what population?

Question 5: Would it be reasonable to conclude that eating more omega–3 fatty acids reduces migraine pain? Explain.

Study 1: Five hundred students were selected at random from those enrolled at large college. Each student in the sample was asked whether they had experienced a migraine in the last week. Each student also completed a survey about their diet, which focused on omega–3 fatty acid consumption. The study found that those whose diets were high in omega–3 fatty acids were significantly less likely to experience migraines.

Study 2: One hundred college students who suffer from frequent migraines volunteered to participate in a statistical study. The students were divided based on whether they live off campus or on campus, since living on campus may limit students' food choices. The students who live off campus were asked to eat at least two servings of omega–3-rich fish per week for a period of 8 weeks. The students who live on campus were asked to continue eating normally. Both groups tracked their migraine headaches over the course of the study, and those who ate the diet rich in omega–3s experienced significantly fewer headache days, on average, than those who ate their regular diets.

Study 3: Two hundred people who suffer from frequent migraines volunteered to participate in a statistical study. Each person in the sample was asked how many migraines they had experienced in the previous week. Each person also completed a survey about their diet, which focused on omega-3 fatty acid consumption. The study found that, on average, those whose diets were high in omega-3 fatty acids had experienced significantly fewer headaches in the previous week.

Study 4: Two hundred people who suffer from frequent migraines volunteered to participate in a statistical study. For each volunteer, a coin was tossed. If the coin landed heads up, the volunteer was assigned to group 1. If the coin landed tails up, the volunteer was assigned to group 2. Those in group 1 were asked to eat at least two servings of omega-3-rich fish per week for a period of 8 weeks. Those in group 2 were asked to continue eating normally. Both groups tracked their migraine headaches over the course of the study, and those who ate the diet rich in omega-3 experienced significantly fewer headache days, on average, than those who ate their regular diets.

Study 5: Five hundred students were selected at random from those enrolled at large college. Each of the selected students was asked to participate in a study, and all agreed to participate. For each volunteer, a coin was tossed. If the coin landed heads up, the volunteer was assigned to group 1. If the coin landed tails up, the volunteer was assigned to group 2. Those in group 1 were asked to eat at least two servings of omega-3-rich fish per week for a period of 8 weeks. Those in group 2 were asked to continue eating normally. At the end of the study, each student was asked whether they had experienced a migraine during the last 8 weeks. The study found that those assigned to eat diets high in omega-3 fatty acids were significantly less likely to experience migraines.

2.7 The article **"Popping Cork Sound Makes Wine Taste Better" (decanter.com, September 28, 2017)** describes a study in which 140 people were assigned to one of two groups. Those in one group heard the noise made by a bottle of wine being uncorked before tasting a glass of wine. Those in the other group heard the noise made by a person releasing a screw cap before tasting a glass of wine. Both groups tasted the same wine. It was reported that the average rating was higher for the group that heard the cork popping than the average rating for the group that heard the screw cap being released.

a. Is the study described an observational study or an experiment?

b. Can a case be made for the researcher's conclusion that hearing a cork pop rather than a screw cap being released was the cause for the higher rating? Explain.

2.8 In a study titled **"Is Screen Time Associated with Anxiety or Depression in Young People?" (*BMC Public Health* [2019]: 82),** researchers assessed associations between the use of screen-based devices at age 16 (measured via questionnaire) and anxiety and depression at age 18 (measured via a self-administered diagnostic instrument called the Revised Clinical Interview Schedule). As part of the study, three populations of teenagers were compared: those who usually spend less than 1 hour on the computer each day, those who usually spend 1 to 2 hours, and those who usually spend 3 or more hours. The study found that participants who spend more time using the computer were slightly more likely to experience symptoms of anxiety.

a. Is this an observational study or an experiment?

b. Is it reasonable to conclude that time spent on the computer is the cause of the observed increase in anxiety? Explain.

2.9 For the study described in the previous exercise, why would it be important to know about how the teens who participated in the study were selected if you wanted to generalize the given conclusion to all teens who spend time on a computer?

2.10 A study described in the online article **"Cheers! Moderate Drinking Linked to Lower Risk of Heart Disease" (studyfinds.org, March 28, 2017)** was an observational study, and the researchers who conducted the study indicated that because it was an observational study, no conclusions can be drawn in terms of cause and effect. What are some potential confounding variables that prevent concluding that moderate drinking is the cause of the lower risk of heart disease?

2.11 A survey of adult Americans carried out in 2021 found that 69% were Facebook users **("Social Media Use in 2021," Pew Research Center, April 7, 2021)**.

a. What condition on how the data were collected would make the generalization from the sample to the population of all adult Americans reasonable?

b. This survey was conducted by telephone. Do you think it would be reasonable to generalize from the sample and say that 69% of all adult Americans use Facebook? Explain.

2.12 Does sitting for long periods of time hurt your heart? The article **"Why Sitting May Be Bad for Your Heart" (*The New York Times*, December 20, 2017)** describes a study of 1700 people who were participants in the Dallas Heart Study. The study found that the people who sat for long periods of time tended to have higher levels of troponin in their blood. Troponin is a protein that is released when the heart muscle has been damaged.

The article states that "Of course, this was an observational study and can show only that sitting is linked to high troponin, not that it causes troponins to rise." Do you agree with this statement? Explain.

Section 2.2 Sampling

Many studies are conducted to learn about a population. In this case, it is important that the sample selected for the study is representative of the population. To be reasonably sure of this, you must think carefully about how the sample is selected.

It is sometimes tempting to take the easy way out and gather data in a haphazard way. But if a sample is chosen on the basis of convenience alone, it is not possible to interpret the resulting data with confidence. For example, it might be easy to use the students in your statistics class as a sample of students at your university. However, not all majors include a statistics course in their curriculum, and most students take statistics in their freshman or sophomore year. If you attempt to generalize from this convenience sample, it is not clear how these factors (and others that you might not be aware of) affect any conclusions based on information from the sample.

There is no way to tell just by looking at a sample whether it is representative of the population from which it was selected. The only assurance comes from the method used to select the sample.

There are many reasons for selecting a sample rather than obtaining information from an entire population (a **census**). Sometimes the process of measuring a characteristic of interest damages the item that is being measured. This is the case with measuring the lifetime of flashlight batteries (which results in dead batteries) or measuring the sugar content of oranges (which results in oranges that cannot be sold). It would be foolish to study the entire population in situations like these. But the most common reason for selecting a sample is limited resources. Restrictions on available time or money usually make it impossible to collect data from an entire population.

What Do You Think?

What is another example of a situation where you would want to use data from a sample rather than using a census of the entire population?

Bias in Sampling

Bias in sampling is the tendency for samples to differ from the corresponding population in some systematic way. Bias can result from the way in which the sample is selected or from the way in which information is obtained once the sample has been chosen. The most common types of bias encountered in sampling situations are selection bias, measurement or response bias, and nonresponse bias.

Selection bias (sometimes also called undercoverage) is introduced when the way the sample is selected systematically excludes some part of the population of interest. For example, a researcher may wish to generalize from the results of a study to the population consisting of all residents of a particular city, but the method of selecting individuals may tend to exclude people who are homeless or without telephones.

If those who are excluded from the sampling process differ in some systematic way from those who are included, the sample will almost always be unrepresentative of the population. If this difference between the included and the excluded occurs on a variable that is important to the study, conclusions based on the sample data may not be valid for the population of interest.

Selection bias also occurs if only volunteers or self-selected individuals are used in a study, because those who choose to participate (for example, in an online poll) may differ from those who choose not to participate.

Measurement or response bias occurs when the method of observation tends to produce values that systematically differ from the true value in some way. This might happen if an improperly calibrated scale is used to weigh items or if questions on a survey are worded in a way that tends to influence the response.

For example, David Wilson in his blog for ***The Huffington Post*** **("Fox News Poll: What Does Bad Question Wording Say About Their Polling Data?" September 21, 2010, huffingtonpost.com)** writes:

> One of the main sources of error in interpreting poll results is poor question wording. In surveys, questions are supposed to be accurate measures of one's attitudes, opinions, beliefs, and behaviors; just like the scales in bathrooms accurately measure one's weight. They should at least be well written and easily understood by respondents, and they should not be biased toward a particular viewpoint.

He goes on to give several examples of "bad" questions, including the following question, which was part of a Fox News poll conducted in September 2010:

> Do you agree or disagree with the following statement: the federal government has gotten totally out of control and threatens our basic liberties unless we clean house and commit to drastic change?

Wilson points out that this question makes a number of assertions, such as the government is out of control and basic liberties are threatened. If someone responded "agree" to this question, it would be unclear exactly which assertion they were in agreement with. In addition, words like *control* and *threaten* may lead people to respond in a particular way.

Other things that might contribute to response bias are the appearance or behavior of the person asking the question, the group or organization conducting the study, and the tendency for people not to be completely honest when asked about illegal behavior or unpopular beliefs.

Although the terms *measurement bias* and *response bias* are often used interchangeably, the term *measurement bias* is usually used to describe systematic deviation from the true value as a result of a faulty measurement instrument. Response bias is typically used to describe systematic deviations from the true value when people provide answers to survey questions.

What Do You Think?

Would you be worried about bias if a survey included the following question? If so, what type of bias?

Question: How would you rate our extremely hard-working customer service team?

Nonresponse bias occurs when responses are not obtained from all individuals selected for inclusion in the sample. As with selection bias, nonresponse bias can distort results if those who respond differ in important ways from those who do not respond. Although some level of nonresponse is unavoidable in most surveys, the biasing effect on the resulting sample is lowest when the response rate is high. To minimize nonresponse bias, it is important that a serious effort be made to follow up with individuals who do not respond to an initial request for information.

The nonresponse rate for surveys or opinion polls varies dramatically, depending on how the data are collected. Surveys are commonly conducted by mail, by phone, and by personal interview. Mail surveys are inexpensive but often have high nonresponse rates. Telephone surveys can also be inexpensive and can be implemented quickly, but they work well only for short surveys and they can also have high nonresponse rates. Personal interviews are generally expensive but tend to have better response rates. Some of the many challenges of conducting surveys are discussed in Section 2.6 (available online).

Types of Bias

Selection Bias

Tendency for samples to differ from the population as a result of systematic exclusion of some part of the population.

Measurement or Response Bias

Tendency for samples to differ from the population because the method of observation tends to produce values that differ from the true values.

Nonresponse Bias

Tendency for samples to differ from the population because data are not obtained from all individuals selected for inclusion in the sample.

It is important to understand that bias is introduced by the way a sample is selected or by the way the data are collected from the sample. Increasing the size of the sample, although possibly desirable for other reasons, does not reduce bias if the method of selecting the sample is flawed or if the nonresponse rate remains high.

Potential sources of bias are illustrated in the following examples.

Example 2.1 Are Cell Phone Users Different?

Understand the context ➤

Many surveys are conducted by telephone and participants are often selected from phone directories that include only landline telephones. For many years, it was thought that this was not a serious problem because most cell phone users also had a landline phone and so they still had a chance of being included in the survey. But the number of people with only cell phones is growing. This trend is a concern for survey organizations, and those who conduct telephone surveys need to worry about how selection bias might influence the ability to generalize survey results if only landlines are used.

Call screening is another concern with telephone surveys described in the article **"Who in the World Is Still Answering Pollsters' Phone Calls?" (*The New York Times*, October 12, 2022).** Many people now only answer the phone when they recognize the phone number of the caller. The article notes that in a telephone poll conducted in 2022, only 0.4% of telephone numbers called resulted in a completed interview. The author of the article wonders if we are seeing the "death of telephone polling." This lack of response raises the possibility of nonresponse bias in the survey results.

Example 2.2 Think before You Order That Burger!

Understand the context ➤

The article **"What People Buy from Fast-Food Restaurants: Caloric Content and Menu Item Selection" (*Obesity* [2009]: 1369–1374)** reported that the average number of calories consumed at lunch in New York City fast-food restaurants was 827. The researchers selected 267 fast-food locations at random. The paper states that at each of these locations "adult customers were approached as they entered the restaurant and asked to provide their food receipt when exiting and to complete a brief survey."

Approaching customers as they entered the restaurant and before they ordered may have influenced what they purchased. This introduces the potential for response bias.

In addition, some people chose not to participate when approached. If those who chose not to participate differed from those who did participate, the researchers also need to be concerned about nonresponse bias. Both of these potential sources of bias limit the researchers' ability to generalize conclusions based on data from this study.

Random Sampling

Many methods introduced in this text rely on random selection. The most straightforward sampling method based on random selection is called simple random sampling. A **simple random sample** is a sample chosen using a method that ensures that each different possible sample of the desired size has an equal chance of being the one chosen.

For example, suppose that you want a simple random sample of 10 employees chosen from all those who work at a large design firm. For the sample to be a simple random sample, the method used to select the sample must ensure that each of the many different subsets of 10 employees are equally likely to be selected. A sample taken from only full-time employees would not be a simple random sample of *all* employees, because someone who works part-time has no chance of being selected. Although a simple random sample may, by chance, include only full-time employees, it must be selected in such a way that each possible sample, and therefore *every* employee, has the same chance of being selected.

It is the selection process, not the final sample, which determines whether a sample is a simple random sample.

The letter n is used to denote sample size. It is the number of individuals or objects in the sample. For the design firm scenario just described, $n = 10$ because 10 employees were to be selected.

Definition

Simple random sample of size n: A sample that is selected from a population in a way that ensures that every different possible sample of size n has the same chance of being selected.

The definition of a simple random sample implies that every individual member of the population has an equal chance of being selected. However, the fact that every individual has an equal chance of selection, *by itself,* is not enough to guarantee that the sample is a simple random sample.

For example, suppose that a class is made up of 100 students, 60 of whom live on campus. A researcher decides to select 6 of the students who live on campus by writing all 60 names on slips of paper, mixing the slips, and then picking 6. They then select 4 students who do not live on campus from the class using a similar procedure. Even though every student in the class has an equal chance of being included in the sample (6 of 60 students who live on campus are selected and 4 of 40 students who do not live on campus are chosen), the resulting sample is *not* a simple random sample because not all different possible samples of 10 students from the class have the same chance of selection. Many possible samples of 10 students—for example, a sample of 7 who live on campus and 3 who do not live on campus or a sample of students who all live on campus—have no chance of being selected. The sample selection method described here is not necessarily a bad choice (in fact, it is an example of stratified sampling, to be discussed in more detail shortly). But it does not produce a simple random sample and different methods would be used to generalize results from the sample to the population. For this reason, the sampling method is an important consideration when a method is chosen for analyzing data.

Selecting a Simple Random Sample

A number of different methods can be used to select a simple random sample. One way is to put the name or number of each member of the population on different but identical slips of paper. The process of thoroughly mixing the slips and then selecting n slips yields a random sample of size n. This method is easy to understand, but it has obvious drawbacks. The mixing must be adequate, and producing the necessary slips of paper can be extremely tedious, even for relatively small populations.

A commonly used method for selecting a random sample is to first create a list, called a **sampling frame**, of the objects or individuals in the population. Each item on the list is identified by a number. A table of random digits or a random number generator can then be used to select the sample. A random number generator is an algorithm that produces a sequence of numbers that satisfies properties associated with the notion of randomness. Most statistics software packages and many calculators include a random number generator. A small table of random digits can be found in Appendix A, Table 1.

For example, suppose a list containing the names of the 427 customers who purchased a new car during 2023 at a car dealership is available. The owner of the dealership wants to interview a sample of these customers to learn about customer satisfaction. They plan to select a simple random sample of 20 customers. Because it would be tedious to write all 427 names on slips of paper, random numbers can be used to select the sample. To do this, three-digit numbers, starting with 001 and ending with 427, can be used to represent the individuals on the list.

The random digits from rows 6 and 7 of Appendix A, Table 1 are shown here:

0 9 3 8 7 6 7 9 9 5 6 2 5 6 5 8 4 2 6 4
4 1 0 1 0 2 2 0 4 7 5 1 1 9 4 7 9 7 5 1

Blocks of three digits from this list (underlined in the lists above) can be used to identify the individuals who should be included in the sample. The first block of three digits is 093, so the 93rd person on the list will be included in the sample. The next five blocks of three digits (876, 799, 562, 565, and 842) do not correspond to anyone on the list, so you ignore them. The next block that corresponds to a person on the list is 410, so that person is included in the sample. This process would continue until 20 people have been selected for the sample. You would ignore any three-digit repeats since any particular person should only be selected once for the sample.

Another way to select the sample would be to use computer software or a graphing calculator to generate 20 random numbers. For example, Minitab produced the following numbers when 20 random numbers between 1 and 427 were requested.

```
289   67   29   26  205  214  422   31  233   98
 10  203  346  186  232  410   43  293   25  371
```

These numbers could be used to determine which 20 customers to include in the sample.

When selecting a random sample, researchers can choose to do the sampling with or without replacement. **Sampling with replacement** means that after each item is selected for the sample, the item is “replaced” back into the population and may therefore be selected again at a later stage. In practice, sampling with replacement is rarely used. Instead, the more common method is to not allow the same item to be included in the sample more than once. After being included in the sample, an individual or object would not be considered for further selection. This is called **sampling without replacement**.

Definitions

Sampling without replacement: Once an individual from the population is selected for inclusion in the sample, it may not be selected again in the sampling process. A sample selected without replacement includes n different individuals from the population.

Copyright 2025 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s). Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

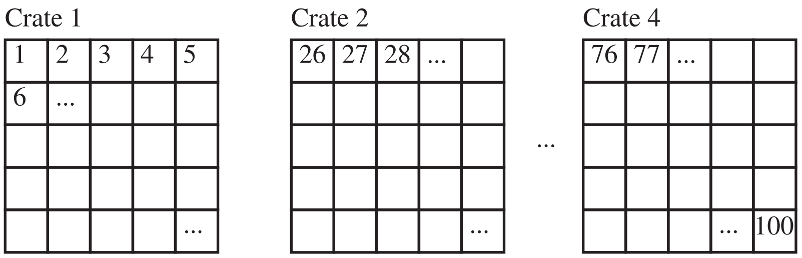
Sampling with replacement: After an individual from the population is selected for inclusion in the sample and the corresponding data are recorded, the individual is placed back in the population and can be selected again in the sampling process. A sample selected with replacement might include any particular individual from the population more than once.

Although these two forms of sampling are different, when the sample size *n* is small relative to the population size, as is often the case, there is little practical difference between them. In practice, the two methods can be viewed as equivalent if the sample size is less than 10% of the population size.

Example 2.3 Selecting a Random Sample of Glass Soda Bottles

Understand the context ➤

Breaking strength is an important characteristic of glass soda bottles. Suppose that you want to measure the breaking strength of each bottle in a random sample of size $n = 3$ selected from four crates containing a total of 100 bottles (the population). Each crate contains five rows of five bottles each. You can identify each bottle with a number from 1 to 100 by numbering across the rows in each crate, starting with the top row of crate 1, as pictured:



Formulate a plan ➤

Using a random number generator from a calculator or statistical software package, you could generate three random numbers between 1 and 100 to determine which bottles would be included in the sample. This might result in bottles 15 (row 3 column 5 of crate 1), 89 (row 3 column 4 of crate 4), and 60 (row 2 column 5 of crate 3) being selected.

The goal of random sampling is to produce a sample that is likely to be representative of the population. Although random sampling does not *guarantee* that the sample will be representative, it does allow us to assess the risk of an unrepresentative sample. It is the ability to quantify this risk that enables you to generalize with confidence from a random sample to the corresponding population.

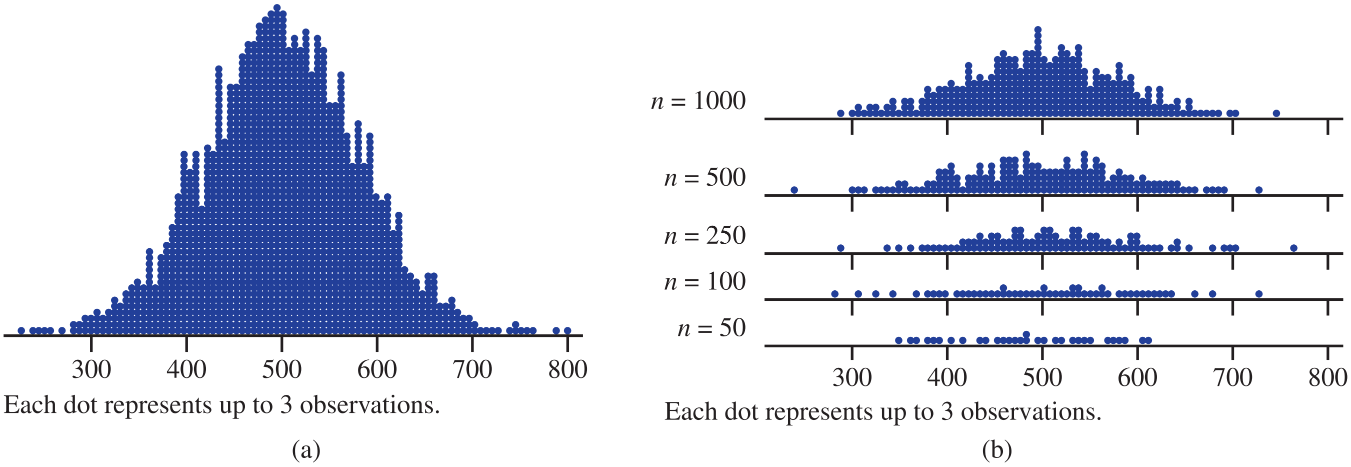
An Important Note Concerning Sample Size

It is a common misconception that if the size of a sample is relatively small compared to the population size, the sample cannot possibly accurately reflect the population. Critics of polls often make statements such as, “There are 21.9 million registered voters in California. How can a sample of 1000 registered voters possibly reflect public opinion when only about 1 in every 22,000 people is included in the sample?” These critics do not understand the power of random selection!

Consider a population consisting of 5000 applicants to a state university, and suppose that you are interested in math SAT scores for this population. A dotplot of the values in this population is shown in Figure 2.1(a). Figure 2.1(b) shows dotplots of the math SAT scores for individuals in five different random samples from the population, ranging in sample size from $n = 50$ to $n = 1000$.

Figure 2.1

(a) Dotplot of math SAT scores for the entire population. (b) Dotplots of math SAT scores for random samples of sizes 50, 100, 250, 500, and 1000.



Notice that each of the samples tends to reflect the distribution of scores in the population. If you were interested in using the sample to estimate the population average or to say something about the variability in math SAT scores, even the smallest of the samples pictured ($n = 50$) would provide reliable information.

Although it is possible to obtain a simple random sample that does not do a reasonable job of representing the population, this is likely to happen only when the sample size is very small. Unless the population itself is small, this risk does not depend on what fraction of the population is sampled. The random selection process allows you to be confident that the resulting sample adequately reflects the population, even when the sample consists of only a small fraction of the population.

Other Sampling Methods

Simple random sampling provides researchers with a sampling method that is objective and free of selection bias. In some settings, however, alternative sampling methods may be less costly, easier to implement, and sometimes even more accurate.

Stratified Random Sampling

When the entire population can be divided into a set of nonoverlapping subgroups, a method known as **stratified sampling** is often easier to implement and more cost-effective than simple random sampling. In stratified random sampling, separate simple random samples are independently selected from each subgroup.

For example, to estimate the average cost of malpractice insurance, a researcher might find it convenient to view the population of all doctors practicing in a particular city as being made up of four groups: (1) surgeons, (2) internists and family practitioners, (3) obstetricians, and (4) a group that includes all other areas of specialization. Rather than taking a simple random sample from the population of all doctors, the researcher could take four separate simple random samples—one from the group of surgeons, another from the internists and family practitioners, and so on. These four samples would provide information about the four subgroups as well as information about the overall population of doctors.

When the population is divided in this way, the subgroups are called **strata** and each individual subgroup is called a stratum (the singular of strata). Stratified sampling selects a separate simple random sample from each stratum. Stratified sampling can be used instead of simple random sampling if it is important to obtain information about characteristics of the individual strata as well as of the entire population, although a stratified sample is not required to do this—subgroup estimates can also be obtained by using an appropriate subset of data from a simple random sample.

Copyright 2025 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s). Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

The real advantage of stratified sampling is that it often allows you to make more accurate inferences about a population than does simple random sampling. In general, it is much easier to produce relatively accurate estimates of characteristics of a homogeneous group than of a heterogeneous group. For example, even with a small sample, it is possible to obtain an accurate estimate of the average grade point average (GPA) of students graduating with high honors from a university. The individual GPAs of these students are all quite similar (a homogeneous group), and even a sample of three or four individuals from this group should be representative. On the other hand, producing a reasonably accurate estimate of the average GPA of *all* seniors at the university, a much more diverse group of GPAs, is a more difficult task. This means that if a varied population can be divided into strata, with each stratum being much more homogeneous than the population with respect to the characteristic of interest, then a stratified random sample can produce more accurate estimates of population characteristics than a simple random sample of the same size.

What Do You Think?

If you were planning to choose strata that would be used to select a stratified random sample, it would not be a good idea to create the strata using random selection from the population. Why not?

Cluster Sampling

Sometimes it is easier to select groups of individuals from a population than it is to select individuals themselves. **Cluster sampling** involves dividing the population of interest into nonoverlapping subgroups, called **clusters**. Clusters are then selected at random, and then *all* individuals in the selected clusters are included in the sample.

For example, suppose that a large urban high school has 600 senior students, all of whom are enrolled in a first period homeroom. There are 24 senior homerooms, each with approximately 25 students. If school administrators wanted to select a sample of about 75 seniors to participate in an evaluation of the college and career placement advising available to students, they might find it much easier to select three of the senior homerooms at random and then include all the students in the selected homerooms in the sample. Then a survey could be administered to all students in the selected homerooms at the same time—certainly easier to implement than randomly selecting 75 individual seniors and then administering the survey to those students.

Because whole clusters are selected in cluster sampling, the ideal situation is when each cluster mirrors the characteristics of the population. When this is the case, a small number of clusters results in a sample that is representative of the population. If it is not reasonable to think that the variability present in the population is reflected in each cluster, as is often the case when the cluster sizes are small, then it becomes important that a large number of clusters are included in the sample.

Be careful not to confuse clustering and stratification. Even though both of these sampling strategies involve dividing the population into subgroups, both the way in which the subgroups are sampled and the best strategy for creating the subgroups are different.

In stratified sampling, you sample from every subgroup, whereas in cluster sampling, you include only selected whole clusters in the sample. Because of this difference, to increase the chance of obtaining a sample that is representative of the population, you want to create homogeneous groups for strata and heterogeneous (reflecting the variability in the population) groups for clusters.

Copyright 2025 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s). Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

What Do You Think?

A large manufacturing plant operates three shifts a day (day, evening, and night). A survey to assess job satisfaction is to be conducted by selecting random samples of 50 employees working each shift. Is this sample selection method stratified random sampling or cluster sampling?

Systematic Sampling

Systematic sampling is a method that can be used when it is possible to view the population of interest as consisting of a list or some other sequential arrangement. A value k is specified (for example, $k = 50$ or $k = 200$). Then one of the first k individuals is selected at random, after which every kth individual in the sequence is included in the sample. A sample selected in this way is called a **1 in *k* systematic sample**.

For example, a sample of faculty members at a university might be selected from the faculty phone directory. One of the first $k = 20$ faculty members listed could be selected at random, and then every 20th faculty member after that on the list would also be included in the sample. This would result in a 1 in 20 systematic sample.

The value of k for a 1 in k systematic sample is generally chosen to achieve a desired sample size. For example, in the faculty directory scenario just described, if there were 900 faculty members at the university, the 1 in 20 systematic sample described would result in a sample size of 45. If a sample size of 100 was desired, a 1 in 9 systematic sample could be used (because $900/100 = 9$).

As long as there are no repeating patterns in the population sequence, systematic sampling usually works reasonably well. However, if there are repeating patterns, systematic sampling can result in an unrepresentative sample. For example, suppose that workers at the entry station of a state park have recorded the number of visitors to the park each day for the past 10 years. In a 1 in 70 systematic sample of days from this list, you would pick 1 of the first 70 days at random and then every 70th day after that. But if the first day selected happened to be a Wednesday, *every* day selected in the entire sample would also be a Wednesday (because there are 7 days a week and 70 is a multiple of 7). It is unlikely that such a sample would be representative of the entire collection of days. The number of visitors is likely to be higher on weekend days, and no Saturdays or Sundays would be included in the sample.

Convenience Sampling: Don't Go There!

It is often tempting to resort to **convenience sampling**—that is, using an easily available or convenient group to form a sample. This is a recipe for disaster! Results from such samples are rarely informative, and it is risky to try to generalize from a convenience sample to any larger population.

One common form of convenience sampling is sometimes called **voluntary response sampling**. Such samples rely entirely on individuals who volunteer to be a part of the sample, often by responding to an advertisement, calling a publicized telephone number to register an opinion, or logging on to an Internet site to complete a survey. It is extremely unlikely that individuals participating in such voluntary response surveys are representative of any larger population of interest.

Exercises 2.13–2.34

2.13 A New York psychologist recommends that if you feel the need to check your cell phone in the middle of a movie or if you sleep with your cell phone next to your bed, it might be time to "power off" (***AARP Bulletin*, September 2010**). Suppose that you want to learn about the proportion of students at your college who would feel the need to check their cell phone during the middle of a movie and that you have access to a list of all students enrolled at your college. Describe how you would use this list to select a simple random sample of 100 students.

Copyright 2025 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s). Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

2.14 As part of a curriculum review, a psychology department would like to select a simple random sample of 20 of last year's 140 graduates to obtain information on how graduates perceived the value of the curriculum. Describe two different methods that might be used to select the sample.

2.15 A petition with 500 signatures is submitted to a university's student council. The council president would like to determine the proportion of those who signed the petition who are actually registered students at the university. There is not enough time to check all 500 names with the registrar, so the council president decides to select a simple random sample of 30 signatures. Describe how this might be done.

2.16 The article **"Bicyclists and Other Cyclists" (*Annals of Emergency Medicine* [2010]: 426)** reported that in 2008, there were 716 bicyclists killed on public roadways in the United States, and that the average age of the cyclists killed was 41 years. These figures were based on an analysis of the records of all traffic-related deaths of bicyclists on U.S. public roadways (this information is kept by the National Highway Traffic Safety Administration).

a. Does the group of 716 bicycle fatalities represent a census or a sample of the 2008 bicycle fatalities?

b. If the population of interest is 2008 bicycle traffic fatalities, is the given average age of 41 years a number that describes a sample or a number that describes the population?

2.17 A survey of Arizona drivers is described in the article **"Study Claims Safety Should Be Made Law" (*Red Rock News*, August 21, 2015).** The following statement is from the article:

> According to the annual survey, which aims to gauge the opinions and concerns of the motoring public across the state regarding traffic safety, Arizona drivers want better traffic safety laws. The survey showed that nine in 10 Arizonans—91 percent—favor a statewide ban on texting while driving for all drivers, and about two-thirds—64 percent—favor a primary seat belt law.

The article also describes how the data for this survey were collected. A survey was mailed to 2500 randomly selected AAA Arizona members (AAA Arizona is an automobile club that provides services like roadside assistance to approximately 860,000 members in Arizona). The article did not indicate how many of the surveys were actually returned.

The results of this survey were generalized to all Arizona drivers. Answer the following four questions for this observational study.

a. What is the population of interest?

b. Was the sample selected in a reasonable way?

c. Is the sample likely to be representative of the population of interest?

d. Are there any obvious sources of bias?

2.18 For each of the situations described, state whether the sampling procedure is simple random sampling, stratified random sampling, cluster sampling, systematic sampling, or convenience sampling.

a. All first-year students at a university are enrolled in one of 30 sections of a seminar course. To select a sample of first-year students at this university, a researcher selects four sections of the seminar course at random from the 30 sections and all students in the four selected sections are included in the sample.

b. To obtain a sample of students, faculty, and staff at a university, a researcher randomly selects 50 faculty members from a list of faculty, 100 students from a list of students, and 30 staff members from a list of staff.

c. A university researcher obtains a sample of students at their university by using the 85 students enrolled in their Psychology 101 class.

d. To obtain a sample of the seniors at a particular high school, a researcher writes the name of each senior on a slip of paper, places the slips in a box and mixes them, and then selects 10 slips. The students whose names are on the selected slips of paper are included in the sample.

e. To obtain a sample of those attending a basketball game, a researcher selects the 24th person through the door. Then, every 50th person after that is also included in the sample.

2.19 Of the 6500 students enrolled at a community college, 3000 are part time and the other 3500 are full time. The college can provide a list of students that is sorted so that all full-time students are listed first, followed by the part-time students.

a. Describe a procedure for selecting a stratified random sample that uses full-time and part-time students as the two strata and that includes 10 students from each stratum.

b. Does every student at this community college have the same chance of being selected for inclusion in this stratified random sample? Explain.

2.20 Briefly explain why it is advisable to avoid the use of convenience samples.

2.21 A sample of pages from this book is to be selected, and the number of words on each page in the sample will be determined. For the purposes of this exercise, equations are not counted as words and a number is counted as a word only if it is spelled out—that is, *ten* is counted as a word, but *10* is not.

a. Describe a sampling procedure that would result in a simple random sample of pages from this book.

b. Describe a sampling procedure that would result in a stratified random sample. Explain why you chose the specific strata used in your sampling plan.

c. Describe a sampling procedure that would result in a systematic sample.

d. Describe a sampling procedure that would result in a cluster sample.

2.22 Using the process you gave in Part (a) of Exercise 2.21, select a simple random sample of at least 20 pages, and record the number of words on each of the selected pages. Construct a dotplot of the resulting sample values, and write a sentence or two commenting on what it reveals about the number of words on a page.

2.23 Using the process you gave in Part (b) of Exercise 2.21, select a stratified random sample that includes a total of at least 20 selected pages, and record the number of words on each of the selected pages. Construct a dotplot of the resulting sample values, and write a sentence or two commenting on what it reveals about the number of words on a page.

2.24 The chairman of a California ballot initiative campaign to add "none of the above" to the list of ballot options in all candidate races was quite critical of a Field poll that showed his measure trailing by 10 percentage points. The poll was based on a random sample of 1000 registered voters in California. They are quoted by the **Associated Press (January 30, 2000)** as saying, "Field's sample in that poll equates to one out of 17,505 voters," and they added that this was so dishonest that Field should get out of the polling business! If you worked on the Field poll, how would you respond to this criticism? (Hint: See discussion of sample size on page 39.)

2.25 The article **"Sleepless in School" (*Journal of American College Health* [2020]: 185–191)** describes a study of 282 college students. The students were recruited from a general education psychology course (a course taken by students from different majors) and they were offered extra credit to fill out an online survey. The survey included questions about the students' sleep health (e.g., hours of sleep during a usual school week) as well as questions about possible social determinants of deficient sleep (e.g., food insecurity, psychological distress).

Answer the following four questions for this observational study.

a. What is the population of interest?

b. Was the sample selected in a reasonable way?

c. Is the sample likely to be representative of the population of interest?

d. Are there any obvious sources of bias?

2.26 The **2013 National Study of Substance Use Habits of College Student-Athletes** surveyed student athletes at NCAA member colleges and universities. The passage below is from the survey website **(ncaa.org)**.

> All NCAA member institutions are asked to participate. The sampling plan achieves an appropriate representation of all NCAA student-athletes while minimizing burden to institutions by asking that all student-athletes on no more than three teams be surveyed on any campus. The teams surveyed are determined by a computer-generated random draw.

The survey is administered to the selected teams in a classroom setting, and no identifying information about the athletes or the college is collected. The website also states

> It is important to note that even with measures to ensure anonymity, self-reported data of this kind can be problematic due to the sensitive nature of the issues. Therefore, absolute levels of use might be underestimated in a study such as this.

a. Was this sample a simple random sample, a stratified sample, a cluster sample, a systematic sample, or a convenience sample? Explain.

b. Give two reasons why an estimate of the proportion of students who reported using illegal drugs based on data from this survey should not be generalized to all U.S. college students.

2.27 The paper **"Deception and Design: The Impact of Communication Technology on Lying Behavior" (*Computer-Human Interaction* [2009]: 130–136)** describes an investigation into whether lying is less common in face-to-face communication than in other forms of communication such as phone conversations or e-mail. Participants in this study were 30 students in an upper-division communications course at Cornell University who received course credit for participation. Participants were asked to record all of their social interactions for a week, making note of any lies told.

Based on data from these records, the authors of the paper concluded that students lie more often in phone conversations than in face-to-face conversations and more often in face-to-face conversations than in e-mail.

Discuss the limitations of this study, commenting on the way the sample was selected and potential sources of bias.

Copyright 2025 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s). Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

2.28 The authors of the paper **"Playing by the Rules: Parental Mediation of Video Game Play" (*Journal of Family Issues* [2015]: 1–24)** used data from a sample of parents to investigate the ways in which parents monitor their children's use of video games. The sample of parents consisted of 427 people who responded to a survey conducted on an Amazon-operated marketplace where people can complete surveys in exchange for compensation.

a. Do you think that the sample of parents was selected in a way that makes it reasonable to think it is representative of the population of all parents?

b. Is it reasonable to generalize conclusions based on data from this survey to all parents? Explain why or why not.

2.29 In 2017, the state of Alabama held a special election to the U.S. Senate. Although Republicans are usually heavily favored in Alabama, this race was contentious, because the Republican candidate had been accused of misconduct. Days before the election, ***FiveThirtyEight* (fivethirtyeight.com, December 11, 2017)** reported large discrepancies in the results from different types of polls. For example, "robo-polls," which involve automated calls to landline phones and often have lower response rates, showed the Republican way ahead. At the same time, a Fox News poll that used a traditional, live-caller survey of both landlines and cell phones showed the Democrat ahead. What factors may have contributed to these very different results?

2.30 In 2019, the Economic Policy Institute published a report titled **"The Teacher Shortage Is Real, Large, and Growing, and Worse than We Thought" (www.epi.org)**. Interested in the causes of the teacher shortage, suppose a researcher wants to conduct a follow-up study measuring job satisfaction among teachers and other professionals who work in schools. They start by dividing educational roles into five groups: administrators, experienced teachers, new teachers, teachers' aides, and other school staff. They then randomly select educators from within each of these five groups. Would this sample be a simple random sample, a systematic sample, a stratified sample, or a cluster sample?

2.31 The financial aid advisor of a university plans to use a stratified random sample to estimate the average amount of money that students spend on textbooks each term. For each of the following proposed stratification schemes, discuss whether it would be worthwhile to stratify the university students in this manner. (Hint: Remember that it is desirable to create strata that are homogeneous.)

a. Strata corresponding to class standing (freshman, sophomore, junior, senior, graduate student)

b. Strata corresponding to field of study, using the following categories: engineering, architecture, business, other

c. Strata corresponding to the first letter of the last name: A–E, F–K, etc.

2.32 Suppose that you were asked to help design a survey of adult city residents in order to estimate the proportion who would support a sales tax increase. The plan is to use a stratified random sample, and three stratification schemes have been proposed.

Scheme 1: Stratify adult residents into four strata based on the first letter of their last name (A–G, H–N, O–T, U–Z).

Scheme 2: Stratify adult residents into three strata: college students, nonstudents who work full time, nonstudents who do not work full time.

Scheme 3: Stratify adult residents into five strata by randomly assigning residents into one of the five strata.

Which of the three stratification schemes would be best in this situation? Explain.

2.33 The article **"High Levels of Mercury Are Found in Californians" (*Los Angeles Times*, February 9, 2006)** describes a study in which hair samples were tested for mercury. The hair samples were obtained from more than 6000 people who voluntarily sent hair samples to researchers at Greenpeace and The Sierra Club. The researchers found that nearly one-third of those tested had mercury levels that exceeded the concentration thought to be safe. Is it reasonable to generalize this result to the larger population of adult Americans? Explain why or why not.

2.34 The **SurveyMonkey** blog **(February 11, 2015)** includes an article titled **"5 Common Survey Question Mistakes That'll Ruin Your Data."** Read this short article, which can be found at the following website, and then answer the following questions (https://www.surveymonkey.com/mp/5-common-survey-mistakes-ruin-your-data/):

a. One of the recommendations in the article is "Don't write leading questions." Give an example of a leading question that is different from the two examples given in the article. Explain why you think the question is a leading question and then suggest a better way to word the question.

b. Select one of the other four recommendations and give an example of a bad question related to that recommendation. Then suggest a better way to word the question.

Section 2.3 Simple Comparative Experiments

Sometimes the questions you are trying to answer are about the effect of certain explanatory variables on some response. Such questions are often of the form, "What happens when . . . ?" or "What is the effect of . . . ?" For example, an industrial engineer may be considering two different workstation designs and might want to know whether the choice of design affects work performance. A medical researcher may want to determine how a proposed treatment for a disease compares to a standard treatment. Experiments provide a way to collect data to answer questions like these.

Definitions

Experiment: A study in which one or more explanatory variables are manipulated in order to observe the effect on a response variable.

Explanatory variables: Those variables that have values that are controlled by the experimenter. Explanatory variables are also called **factors**.

Response variable: A variable that is thought to be related to the explanatory variables in an experiment. It is measured as part of the experiment, but it is not controlled by the experimenter.

Experimental condition: Any particular combination of values for the explanatory variables. Experimental conditions are also called **treatments**.

For example, suppose you are interested in determining the effect of room temperature on performance on a first-year calculus exam. In this case, the explanatory variable is room temperature (it can be manipulated by the experimenter). The response variable is exam performance (the variable that is not controlled by the experimenter and that will be measured).

In general, you can identify the explanatory variables and the response variable easily if you can describe the purpose of the experiment in the following terms:

The purpose is to assess the effect of $\underset{\text{explanatory variable}}{\underline{\hspace{3cm}}}$ on $\underset{\text{response variable}}{\underline{\hspace{3cm}}}$.

Let's return to the example of an experiment to assess the effect of room temperature on exam performance. You might decide to use two room temperature settings, 65° and 75°. This would result in an experiment with two experimental conditions (or equivalently, two treatments) corresponding to the two temperature settings.

Suppose that there are 10 sections of first-semester calculus that have agreed to participate in your study. You might design an experiment in this way: Set the room temperature (in degrees Fahrenheit) to 65° in five of the rooms and to 75° in the other five rooms on test day, and then compare the exam scores for the 65° group and the 75° group. Suppose that the average exam score for the students in the 65° group was noticeably higher than the average for the 75° group. Could you conclude that the increased temperature resulted in a lower average score?

Based on the information given, the answer is no because many other factors might be related to exam score. Were the sections at different times of the day? Did they have different instructors? Different textbooks? Did the sections differ with respect to the abilities of the students? Any of these other factors could provide a plausible explanation (having nothing to do with room temperature) for why the average test score was different for the two groups. It is not possible to separate the effect of temperature from the effects of these other variables. As a consequence, simply setting the room temperatures as described results in a poorly designed experiment.

A well-designed experiment requires more than just manipulating the explanatory variables. The design must also eliminate other possible explanations for any observed differences in the response variable.

Copyright 2025 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s). Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

The goal is to design an experiment that will allow you to determine the effects of the explanatory variables on the chosen response variable. To do this, you must take into consideration any **extraneous variables** that, although not of interest in the current study, might also affect the response variable.

Definition

Extraneous variable: A variable that is not one of the explanatory variables in the study but is thought to affect the response variable.

What Do You Think?

If the room temperature experiment was carried out as described above, what is an example of an extraneous variable?

A well-designed experiment copes with the potential effects of extraneous variables by using **random assignment** to experimental conditions and sometimes also by incorporating direct control and/or blocking into the design of the experiment. Each of these strategies—random assignment, direct control, and blocking—is described in the paragraphs that follow.

A researcher can **directly control** some extraneous variables. In the calculus test example, the textbook used is an extraneous variable because part of the differences in test results might be attributed to this variable. You could control this variable directly, by requiring that all sections use the same textbook. Then any observed differences in test scores between temperature groups could not be explained by the use of different textbooks. The extraneous variable *time of day* might also be directly controlled in this way by having all sections meet at the same time.

The effects of some extraneous variables can be taken into account by a process known as **blocking**. Extraneous variables that are addressed through blocking are called *blocking variables*. An investigator using blocking creates groups (called blocks) that are similar with respect to blocking variables. Then all treatments are tried in each block. In the calculus test example, you might use *instructor* as a blocking variable. If five instructors are each teaching two sections of calculus, you would make sure that for each instructor, one section was part of the 65° group and the other section was part of the 75° group. With this design, if you see a difference in exam scores for the two temperature groups, the extraneous variable *instructor* can be ruled out as a possible explanation, because all five instructors' students were present in each temperature group. (Had you controlled the instructor variable by choosing to have only one instructor, that would be an example of direct control. Of course you can't directly control both time of day and instructor.)

If one instructor taught all the 65° sections and another taught all the 75° sections, you would be unable to distinguish the effect of temperature from the effect of the instructor. In this situation, the two variables (temperature and instructor) are said to be **confounded**.

Two variables are **confounded** if their effects on the response variable cannot be distinguished from one another.

What Do You Think?

If all of the 65° sections were taught at 10 A.M. and all of the 75° sections were taught at 3 P.M., would time of day be confounded with temperature?

If an extraneous variable is confounded with the explanatory variables (which define the treatments), it is not possible to draw a clear conclusion about the effect of the treatment on the response. Both direct control and blocking are effective in ensuring that the controlled variables and blocking variables are not confounded with the variables that define the treatments.

You can directly control some extraneous variables by holding them constant, and you can use blocking to create groups that are similar to essentially filter out the effect of other extraneous variables. But what about variables, such as student ability in the calculus test example, which cannot be controlled by the experimenter and which would be difficult to use as blocking variables? These extraneous variables are handled by the use of **random assignment** to experimental groups.

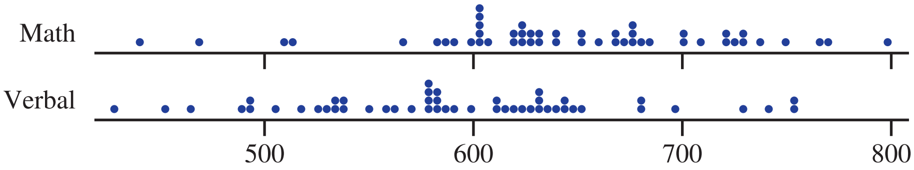
Random assignment ensures that the experiment does not systematically favor one experimental condition over any other and attempts to create experimental groups that are as similar as possible. For example, if the students requesting calculus could be assigned to one of the ten available sections using a random mechanism, you would expect the resulting sections to be similar with respect to student ability as well as with respect to other extraneous variables that are not directly controlled or used as a basis for blocking.

Notice that random assignment in an experiment is different from random selection of subjects. The ideal situation would be to have both random selection of subjects and random assignment of subjects to experimental conditions, as this would allow conclusions from the experiment to be generalized to a larger population.

For many experiments the random selection of subjects is not possible. As long as subjects are assigned at random to experimental conditions, it is still possible to assess treatment effects.

To get a sense of how random assignment tends to create similar groups, suppose that 50 college freshmen are available to participate as subjects in an experiment to investigate whether completing an online review of course material before an exam improves exam performance. The 50 subjects vary quite a bit with respect to achievement, which is reflected in their math and verbal SAT scores, as shown in Figure 2.2.

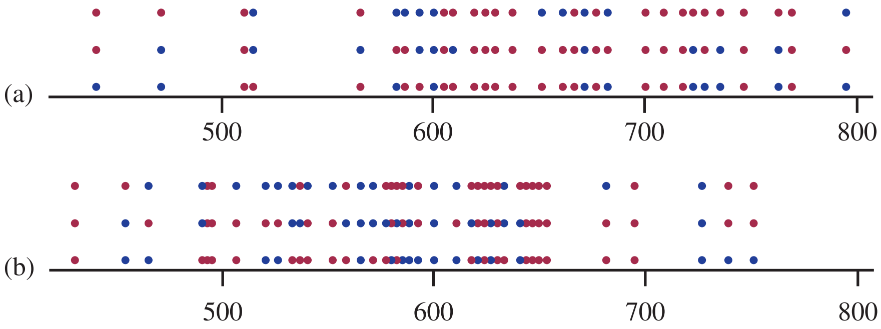
Figure 2.2
Dotplots of math and verbal SAT scores for 50 freshmen.



If these 50 students are to be assigned to the two experimental groups (one that will complete the online review and one that will not), you want to make sure that the assignment of students to groups does not favor one group over the other by tending to assign the higher achieving students to one group and the lower achieving students to the other.

Creating groups of students with similar achievement levels in a way that considers both verbal and math SAT scores simultaneously would be difficult, so you can rely on random assignment. Figure 2.3(a) shows the math SAT scores of the students assigned to each of the two experimental groups (one shown in red and one shown in blue) for each of three different random assignments of students to groups. Figure 2.3(b) shows the verbal SAT scores for the two experimental groups for each of the same three random assignments.

Figure 2.3
Dotplots for three different random assignments to two groups, one shown in red and one shown in blue:
(a) math SAT score;
(b) verbal SAT score.



Copyright 2025 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s). Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

Notice that each of the three random assignments produced groups that are similar with respect to *both* verbal and math SAT scores. So, if any of these three assignments were used and the two groups differed on exam performance, you could rule out differences in math or verbal SAT scores as possible competing explanations for the difference.

Not only will random assignment tend to create groups that are similar with respect to verbal and math SAT scores, but it will also tend to even out the groups with respect to other extraneous variables.

As long as the number of subjects is not too small, you can rely on random assignment to produce comparable experimental groups. This is the reason that random assignment is a part of all well-designed experiments.

Not all experiments require the use of human subjects. For example, a researcher interested in comparing the effect of three different gasoline additives on gas mileage might conduct an experiment using a single car with an empty tank. One gallon of gas with one of the additives will be put in the tank, and the car will be driven along a standard route at a constant speed until it runs out of gas. The total distance traveled on the gallon of gas could then be recorded. This could be repeated a number of times—10, for example—with each additive.

The experiment just described can be viewed as consisting of a sequence of trials. Because a number of extraneous variables (such as variations in environmental conditions like wind speed or humidity and small variations in the condition of the car) might have an effect on gas mileage, it would not be a good idea to use additive 1 for the first 10 trials, additive 2 for the next 10 trials, and so on. A better approach would be to randomly assign additive 1 to 10 of the 30 planned trials, and then randomly assign additive 2 to 10 of the remaining 20 trials. The resulting plan for carrying out the experiment might look as follows:

Trial	1	2	3	4	5	6	7	. . .	30
Additive	2	2	3	3	2	1	2	. . .	1

When an experiment can be viewed as a sequence of trials, random assignment involves the random assignment of treatments to trials. Remember that *random assignment*—either of subjects to treatments or of treatments to trials—*is a critical component* of a good experiment.

Random assignment can be effective only if the number of subjects or observations in each experimental condition (treatment) is large enough for each experimental group to reliably reflect variability in the overall group. For example, if there were only 20 students requesting calculus, it is unlikely that you would get equivalent groups for comparison, even with random assignment to the ten sections. **Replication** is the design strategy of making multiple observations for each experimental condition. Together, replication and random assignment allow the researcher to be reasonably confident of comparable experimental groups.

To illustrate the design of a simple experiment, consider the dilemma of a server in a local restaurant who would like to increase the amount of their tips. The server's strategy is to write "Thank you" on the back of some checks before giving them to the patrons and write nothing on the other checks. They plan to calculate the percentage of the tip as a measure of success (for example, a 15% tip is common). The server will compare the average percentage of the tips calculated from checks with and without the handwritten "Thank you." If writing "Thank you" does not produce higher tips, the server may try a different strategy.

Principles of Experimental Design

Random Assignment

Random assignment (of subjects to treatments or of treatments to trials) to ensure that the experiment does not systematically favor one experimental condition (treatment) over another.

Blocking

Using extraneous variables to create groups (blocks) that are similar. All experimental conditions (treatments) are then tried in each block.

Direct Control

Holding extraneous variables constant so that their effects are not confounded with those of the experimental conditions (treatments).

Replication

Ensuring that there is an adequate number of observations for each experimental condition.

The server wonders if writing "Thank you" on the checks will have an effect on the amount the customer tips. The writing of "Thank you" and the not writing of "Thank you" are the **treatments** (the two experimental conditions to be compared in the experiment). The two treatments together are the possible values of the **explanatory variable**. The tipping percentage is the **response variable**. The idea behind this terminology is that the tipping percentage is a *response* to the treatments *writing "Thank you"* or *not writing "Thank you."*

This experiment may be thought of as an attempt to explain the variability in the response variable in terms of its presumed cause, the variability in the explanatory variable. That is, as the server manipulates the explanatory variable, they expect the customer response to vary. The server has a good start, but now they must consider the four fundamental design principles.

Replication. The server cannot run a successful experiment by gathering tipping information on only one person for each treatment. There is no reason to believe that any single tipping incident is representative of what would happen in other incidents, and therefore it would be impossible to evaluate the two treatments with only two subjects. To interpret the effects of a particular treatment, the server must **replicate** each treatment in the experiment.

Blocking. Suppose that the server works on both Thursdays and Fridays. Because day of the week might affect tipping behavior, they should block on day of the week and make sure that observations for both treatments are made on each of the 2 days.

Direct Control and Random Assignment. There are a number of extraneous variables that might have an effect on the size of tip. Some restaurant patrons will be seated near the window with a nice view. Some will have to wait for a table, whereas others may be seated immediately. Some may be on a fixed income and cannot afford a large tip. Some of these variables can be directly controlled. For example, the server may choose to use only window tables in the experiment, thus eliminating table location as a potential confounding variable. Other variables, such as length of wait and customer income, cannot be easily controlled. As a result, it is important that random assignment is used to decide which of the window tables will be in the "Thank you" group and which will be in the "No thank you" group. The server might do this by flipping a coin as they prepare the check for each window table. If the coin lands with the head side up, they could write "Thank you" on the bill, omitting the "Thank you" when a tail is observed.

The accompanying box summarizes how experimental designs deal with extraneous variables.

Taking Extraneous Variables into Account

Extraneous variables are variables other than the explanatory variables in an experiment that may also have an effect on the response variable. There are several strategies for dealing with extraneous variables in order to avoid confounding.

Extraneous variables that you know about and choose to incorporate into the experimental design:

Strategies

Direct control—holds extraneous variables fixed so that they can't affect the response variable

Blocking—allows for valid comparisons because each treatment is tried in each block

Sometimes extraneous variables are unknown to the experimenter, or might be impossible or very expensive to control.

Extraneous variables that are not incorporated into the experimental design through direct control or blocking:

Strategy

Random assignment

Extraneous variables that are not incorporated into the design of the experiment are sometimes called **lurking variables.**

A Note on Random Assignment

There are several strategies that can be used to perform random assignment of subjects to treatments or treatments to trials. Two common strategies are:

- Write the name of each subject or a unique number that corresponds to a subject on a slip of paper. Place all of the slips in a container and mix well. Then draw out the desired number of slips to determine those that will be assigned to the first treatment group. This process of drawing slips of paper then continues until all treatment groups have been determined.
- Assign each subject a unique number from 1 to *n,* where *n* represents the total number of subjects. Use a random number generator or table of random numbers to obtain numbers that will identify which subjects will be assigned to the first treatment group. This process would be repeated, ignoring any random numbers generated that correspond to subjects that have already been assigned to a treatment group, until all treatment groups have been formed.

The two strategies above work well and can be used for experiments in which the desired number of subjects in each treatment group has been predetermined.

Another strategy that is sometimes employed is to use a random mechanism (such as tossing a coin or rolling a die) to determine which treatment will be assigned to a particular subject. For example, in an experiment with two treatments, you might toss a coin to determine if the first subject is assigned to treatment 1 or treatment 2. This could continue for each subject—if the coin lands H, the subject is assigned to treatment 1, and if the coin lands T, the subject is assigned to treatment 2. This strategy is fine, but may result in treatment groups of unequal size. For example, in an experiment with 100 subjects, 53 might be assigned to treatment 1 and 47 to treatment 2. If this is acceptable, the coin flip strategy is a reasonable way to assign subjects to treatments.

But suppose you want to ensure that there is an equal number of subjects in each treatment group. Is it acceptable to use the coin flip strategy until one treatment group is complete and then just assign all of the remaining subjects to groups that are not yet full? The answer to this question is that it is probably not acceptable. For example, suppose a list of 20 subjects is in order by age from youngest to oldest and that you want to form two treatment groups each consisting of 10 subjects. Tossing a coin to make the assignments might result in the following (based on using the first row of random digits in Appendix A, Table 1, with an even number representing H and an odd number representing T):

Subject	Random Number	Coin Toss Equivalent	Treatment Group
1	4	H	1
2	5	T	2
3	1	T	2
4	8	H	1
5	5	T	2
6	0	H	1
7	3	T	2
8	3	T	2
9	7	T	2
10	1	T	2
11	2	H	1
12	8	H	1
13	4	H	1
14	5	T	2
15	1	T	2
16	Treatment group 2 filled. Assign all others to treatment group 1.		1
17			1
18			1
19			1
20			1

If the list of subjects was ordered by age, treatment group 1 would end up with a disproportionate number of older people. This strategy usually results in one treatment group drawing disproportionately from the end of the list. This means that the only time the strategy of assigning at random until groups fill up and then assigning the remaining subjects to the group that is not full is reasonable is if you can be sure that the list is in random order with respect to all variables that might be related to the response variable. Because of this, it is best to avoid this strategy. Activity 2.5 investigates potential difficulties with this type of strategy.

On the other hand, if the number of subjects is large, it may not be important that every treatment group has exactly the same number of subjects. If this is the case, it is reasonable to use a coin flip strategy (or other strategies of this type) that does not involve stopping assignment of subjects to a group that becomes full.

Evaluating an Experimental Design

The experimental design principles (see page 50) provide a framework for thinking about an experimental design, as illustrated in the following examples.

Example 2.4 Put Away That Laptop!

The article **"The Pen Is Mightier Than the Keyboard" (*Psychological Science* [2014]: 1–10)** describes several experiments designed to investigate whether the method that students use to take notes has an effect on learning. In one of the experiments described in the article, 67 students from Princeton University were assigned to one of two groups.

Both groups were shown a video of a TED Talk (seeted.com/talks) and asked to take notes during the talk. One group watched the talk in a room equipped with laptops that were not connected to the Internet, and they were asked to use the laptops to take notes. The other group watched the talk in a room where they were given notebooks and asked to take their notes by hand.

When the talk was finished, students were engaged in other tasks for 30 minutes and then were asked to take a test on the material from the TED talk. The test included both fact-recall questions and conceptual-application questions.

The researchers found that the two groups performed equally well on the fact-recall questions, but that on the conceptual-application questions, the group that used laptops to take notes performed significantly worse than the group that took notes by hand.

If you assume that the researcher randomly assigned the subjects to the two groups, then this study is an experiment that compares two treatments (laptop used to take notes and handwritten notes). The responses measured were the score on the fact-recall questions part of the test and the score on the conceptual-application questions part of the test. The experiment used replication (many subjects in each treatment group) and random assignment to control for extraneous variables that might affect the response.

Example 2.5 Morality in the Morning

The article **"The Morning Morality Effect: The Influence of Time of Day on Unethical Behavior" (*Psychological Science* [2014]: 95–102)** describes four studies that investigated whether people are more ethical in the morning than in the afternoon. In one of the studies (Experiment 1), volunteers selected either a morning or an afternoon session. During the session, participants were paid to complete a task that presented opportunities to increase the amount earned by purposely providing incorrect responses. The researchers concluded that participants in the afternoon sessions cheated in order to increase the amount they would be paid significantly more often than those in the morning session.

What Do You Think?

Can you spot an obvious flaw in this study?

You probably noticed that the researchers did not randomly assign participants to the two experimental conditions (morning time and afternoon time). This flaw was noted in the paper, and the authors stated

> An important limitation of the two previous experiments was that participants self-selected a morning or afternoon session. It is possible that unethical people, in general, are more likely to sign up for afternoon sessions than ethical people are; if true, this would provide an alternative explanation for our previous findings.

To address this limitation, two other experiments were also carried out. In one of these experiments, 70 volunteers were randomly assigned to a morning or afternoon session. In these sessions, participants were shown 20 sets each containing 12 numbers with three digits, such as 6.38. For each of these 20 sets, participants were instructed to look for a pair of numbers that added to 10 (such as 3.86 and 6.14). Each set of numbers was shown for 15 seconds, and then participants indicated whether or not they had found a pair of numbers that added to 10. Participants did not have to specify what those numbers were, and they were told that they would earn money for each pair that they found. Although the participants did not know this, only 10 of the 20 sets of numbers actually contained two numbers that added to 10. This allowed the researchers to assess whether people were cheating on the task by saying they had found the numbers in order to increase the amount they earned. The researchers found that those in the afternoon session reported finding significantly more pairs that added to 10 than the participants in the morning session. This led researchers to conclude that people are more honest in the morning.

Many experiments compare a group that receives a particular treatment to a **control group** that receives no treatment.

Example 2.6 Chilling Newborns? Then You Need a Control Group . . .

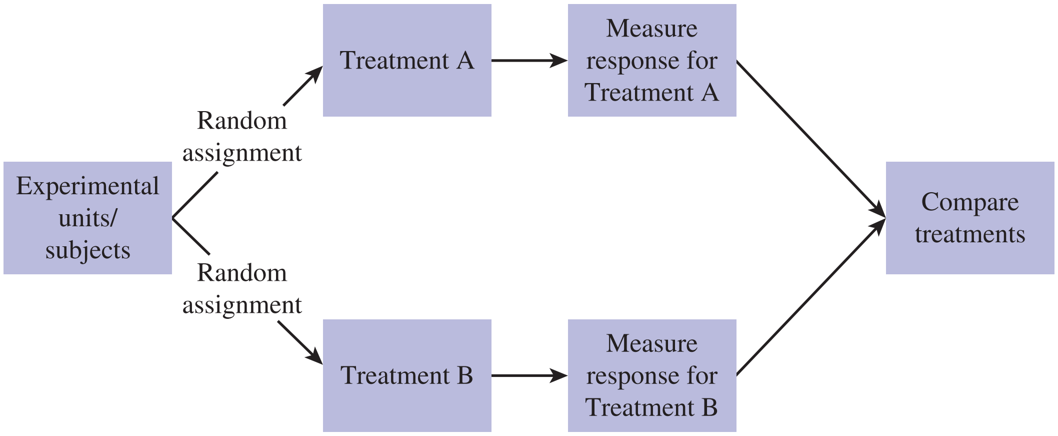
Researchers for the National Institute of Child Health and Human Development studied 208 infants whose brains were temporarily deprived of oxygen as a result of complications at birth (*The New England Journal of Medicine*, October 13, 2005). These babies were subjects in an experiment to determine if reducing body temperature for 3 days after birth improved their chances of surviving without brain damage. The experiment was summarized in a paper that stated "infants were randomly assigned to usual care (control group) or whole-body cooling." Including a control group in the experiment provided a basis for comparison of death and disability rates for the proposed cooling treatment and those for usual care.

Some extraneous variables that might also affect death and disability rates, such as the duration of oxygen deprivation, could not be directly controlled, so to ensure that the experiment did not unintentionally favor one experimental condition over the other, random assignment of the infants to the two groups was critical. Because this was a well-designed experiment, the researchers were able to use the resulting data and statistical methods that you will see in Chapter 11 to conclude that cooling did reduce the risk of death and disability for infants deprived of oxygen at birth.

Visualizing the Underlying Structure of Some Common Experimental Designs

Simple diagrams are sometimes used to highlight important features of experimental designs. The structure of an experiment that is based on random assignment of experimental units (the units to which treatments are assigned, usually subjects or trials) to one of two treatments is displayed in Figure 2.4. The diagram can be easily adapted for an experiment with more than two treatments. In any particular setting, you would also want to customize the diagram by indicating what the treatments are and what response will be measured. This is illustrated in Example 2.7.

Figure 2.4
Diagram of an experiment with random assignment of experimental units to two treatments.


Experimental units/ subjects
Random assignment
Random assignment
Treatment A
Treatment B
Measure response for Treatment A
Measure response for Treatment B
Compare treatments

Example 2.7 Climate Impact Food Labels

Understand the context ➤

Would adding climate impact labels to the food options on a fast food restaurant menu have an effect on what people order at the restaurant? This is the question investigated by the authors of the paper "Effect of Climate Change Impact Menu Labels on Fast Food Ordering Choices Among US Adults" (*JAMA Network Open* [2022]: e2248320). An experiment was

Copyright 2025 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s). Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

conducted in which participants were shown a fast food menu and asked to select an item that they would like to order for dinner. The participants were randomly assigned to one of three groups. There were a total of 5049 participants, with 1666 assigned to group 1, 1667 assigned to group 2, and 1716 assigned to group 3.

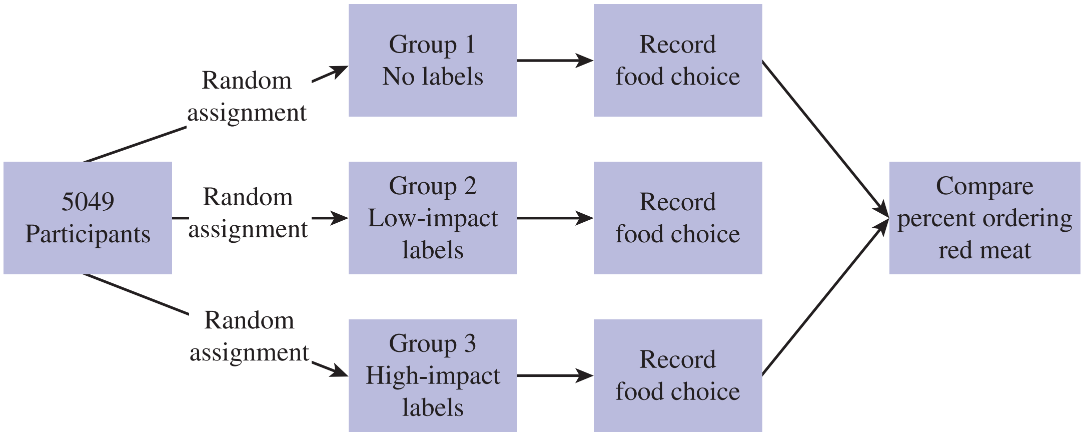
The participants in the first group were given a menu that did not include climate-impact labels. The participants in the second group received menus in which low climate-impact food choices (chicken, fish, and vegetarian items) were labeled with a green low climate-impact label, which was worded as "This item is environmentally sustainable. It has low greenhouse gas emissions and a low contribution to climate change." The third group of participants were given a menu in which high climate-impact items (red meat menu items) were labeled with a red high climate-impact label, which was worded as "This item is not environmentally sustainable. It has high greenhouse gas emissions and a high contribution to climate change." The researchers recorded whether or not the participant ordered a red meat menu item.

What Do You Think?

For this experiment:
What is the explanatory variable?
What are the experimental conditions (treatments)?
What is the response variable?

The researchers used the resulting data to conclude that percentage of participants ordering a red meat item was significantly lower for those who were given a menu with climate-impact labels compared to participants who received a menu without climate labels, and that the difference was greatest for those who received menus with high climate-impact labels. They concluded that climate-impact menu labels, especially high climate-impact labels, may be an effective strategy to promote more sustainable food choices.

Figure 2.5
Diagram for the experiment of Example 2.7.

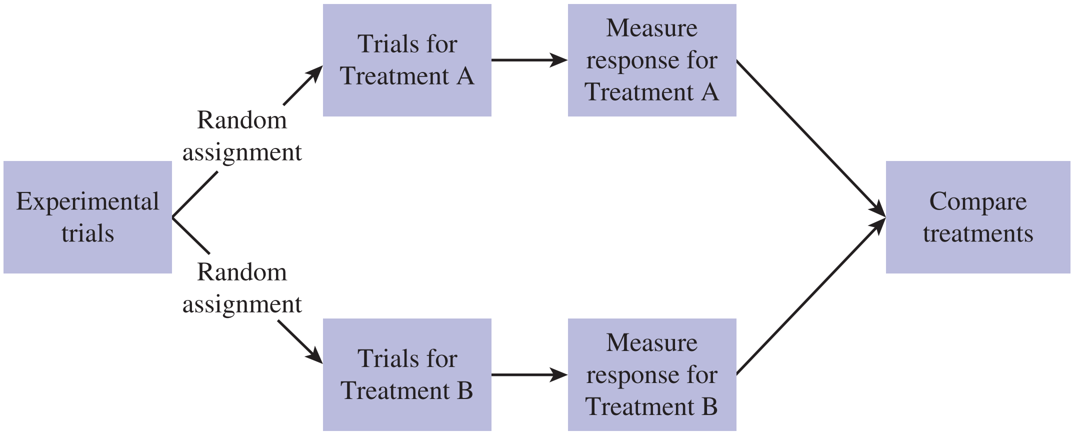


The basic structure of this experiment can be diagramed as shown in Figure 2.5. This type of diagram provides a nice summary of the experiment, but notice that several important characteristics of the experiment are not captured in the diagram. For example, the diagram does not show how participants were selected or if any extraneous variables were directly controlled. This means that while *a diagram* of an experiment may be a useful tool, it usually *cannot stand alone* in describing an experimental design.

Some experiments consist of a sequence of trials, and treatments are assigned at random to the trials. The diagram in Figure 2.6 illustrates the underlying structure of such an experiment. Example 2.8 shows how this diagram can be customized to describe a particular experiment.

Copyright 2025 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s). Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

Figure 2.6
Diagram of an experiment with random assignment of treatments to trials.


Experimental trials
Random assignment
Random assignment
Trials for Treatment A
Trials for Treatment B
Measure response for Treatment A
Measure response for Treatment B
Compare treatments

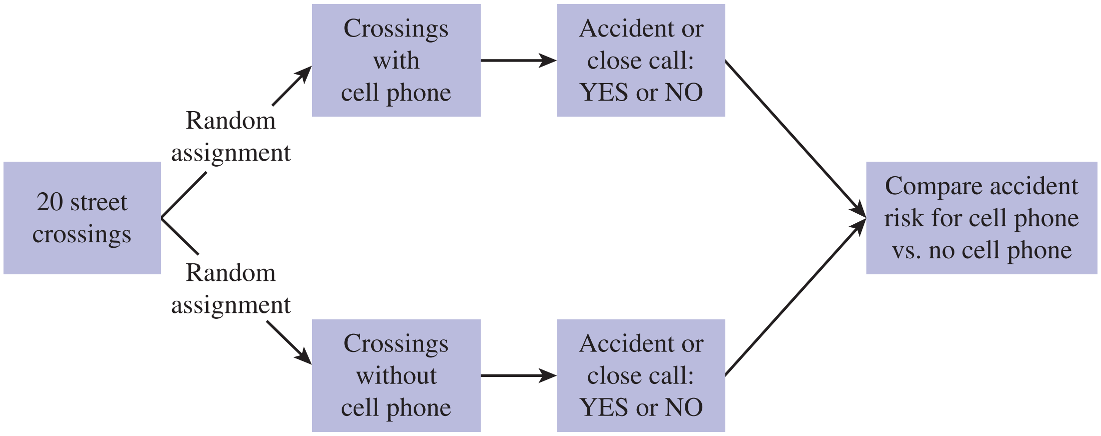
Example 2.8 Distracted? Watch Out for Those Cars!

Understand the context ➤

The paper **"Effect of Cell Phone Distraction on Pediatric Pedestrian Injury Risk" (*Pediatrics* [2009]: e179–e185)** describes an experiment to investigate whether pedestrians who are talking on a cell phone are at greater risk of an accident when crossing the street than when not talking on a cell phone. No children were harmed in this experiment—a virtual interactive pedestrian environment was used! One possible way of conducting such an experiment would be to have a person cross 20 streets in this virtual environment. The person would talk on a cell phone for some crossings and would not use the cell phone for others.

It would be important to randomly assign the two treatments (talking on the phone, not talking on the phone) to the 20 trials (the 20 simulated street crossings). This would result in a design that did not favor one treatment over the other because the pedestrian became more careful with experience or more tired and, therefore, more easily distracted over time. The basic structure of this experiment is diagramed in Figure 2.7.

Figure 2.7
Diagram for the experiment of Example 2.8 with random assignment to trials.

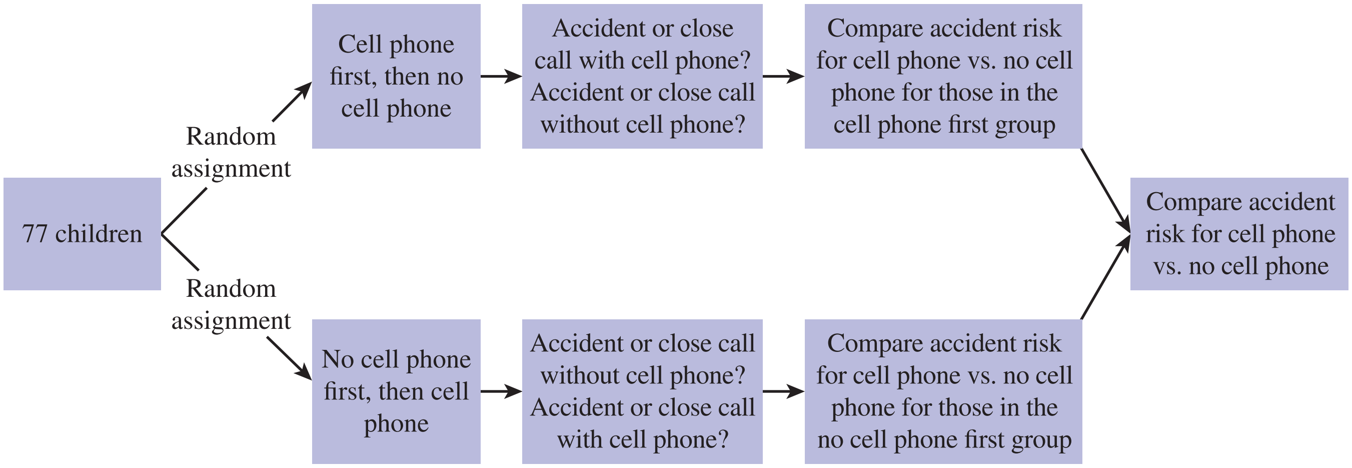

20 street crossings
Random assignment
Random assignment
Crossings with cell phone
Crossings without cell phone
Accident or close call: YES or NO
Accident or close call: YES or NO
Compare accident risk for cell phone vs. no cell phone

The actual experiment conducted by the authors of the paper was a bit more sophisticated than the one just described. In this experiment, 77 children age 10 and 11 each performed simulated crossings with and without a cell phone. Random assignment was used to decide which children would cross first with the cell phone followed by no cell phone and which children could cross first with no cell phone. The structure of this experiment is diagramed in Figure 2.8.

Copyright 2025 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s). Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

Figure 2.8

Diagram for the experiment of Example 2.8 with 77 children.


77 children
Random assignment
Random assignment
Cell phone first, then no cell phone
Accident or close call with cell phone? Accident or close call without cell phone?
Compare accident risk for cell phone vs. no cell phone for those in the cell phone first group
No cell phone first, then cell phone
Accident or close call without cell phone? Accident or close call with cell phone?
Compare accident risk for cell phone vs. no cell phone for those in the no cell phone first group
Compare accident risk for cell phone vs. no cell phone

As was the case in Example 2.7, notice that while the diagram is informative, by itself, it does not capture all of the important aspects of the design. In particular, it does not capture the direct control of age (only children age 10 and 11 were used as subjects in the experiment).

Experimental designs in which experimental units are assigned at random to treatments or in which treatments are assigned at random to trials (like those of the experiments in Examples 2.7 and 2.8) are called **completely randomized designs.**

Diagrams are also useful for highlighting the structure of experiments that use blocking. This is illustrated in Example 2.9.

Example 2.9 Climate-Impact Food Labels Revisited

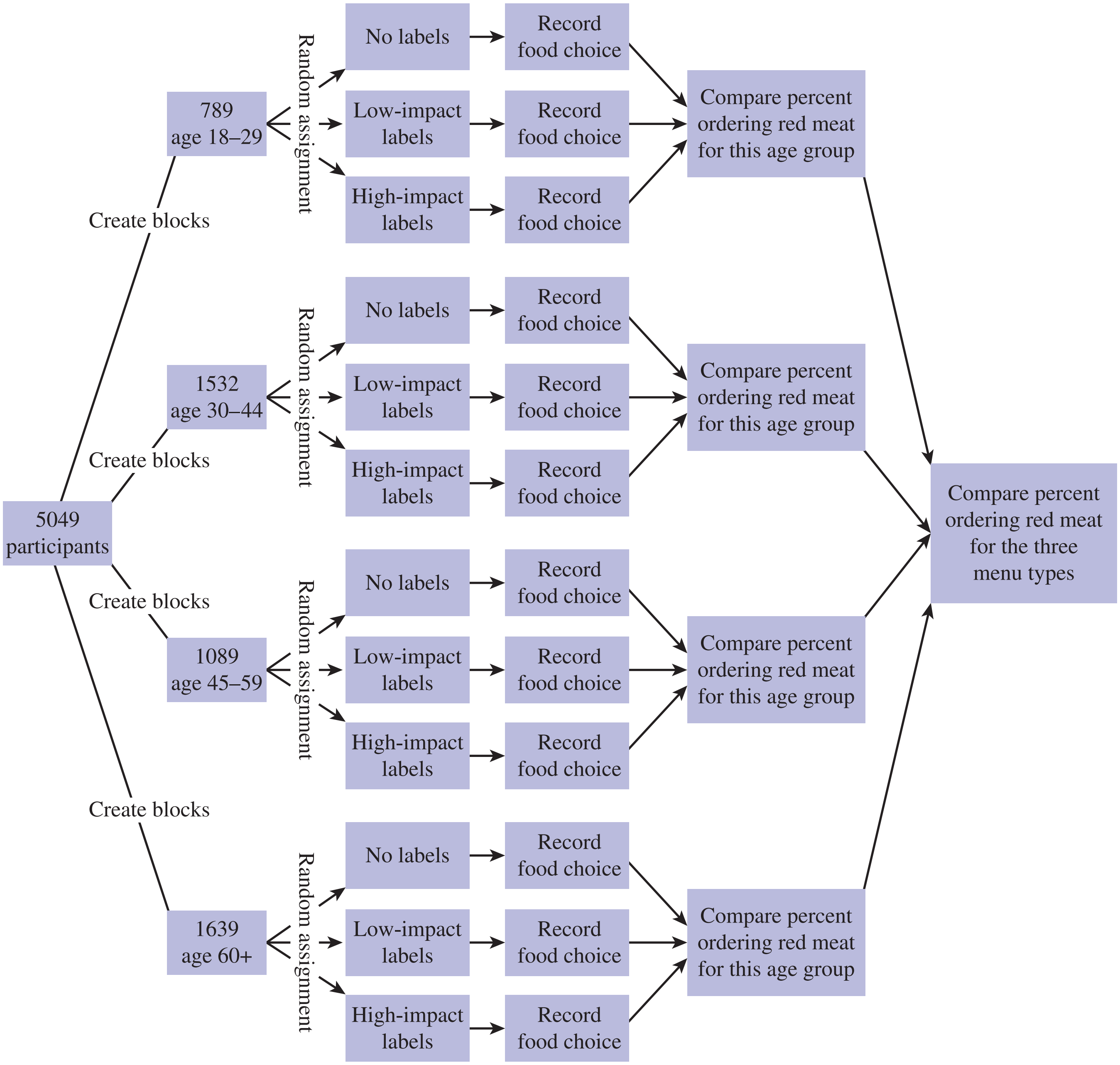
Let’s return to the experiment described in Example 2.7. Take a minute to go back and re-read that example. The experiment described in Example 2.7, a completely randomized design with 5049 participants, was used to compare three different types of fast food menus (no labels, low-impact labels, and high-impact labels). The participants were randomly assigned to the three experimental conditions (treatments) and the researchers relied on random assignment to create groups that would be similar with respect to extraneous variables such as age and ethnicity.

But suppose that you were worried that age might also be related to the type of menu item selected. One possibility would be to use direct control of age—that is you might use only participants who were between 25 and 30 years old. Then if you saw a difference in the percentages who chose a red meat item from the menu, it could not be due to one experimental group including more older people than the other groups. The downside to this strategy is that there is no basis for also generalizing any conclusions from the experiment to other age groups.

Another strategy for dealing with extraneous variables is to incorporate blocking into the design. In the case of age, you might first group the participants based on age to create 4 blocks. For example, the blocks could be constructed by dividing participants into those who were age 18 to 29, those who were age 30 to 44, those who were age 45 to 59, and those who were age 60 and older. Once these blocks were formed, you would randomly assign those in the 18 to 29 age block to the three treatments, and then repeat this process for the other blocks. This would ensure that the three treatment groups were similar with respect to age.

In the actual study, the group of 5049 participants included 789 who were age 18 to 29, 1532 who were age 30 to 44, 1089 who were age 45 to 59, and 1639 who were age 60 and older. A diagram that shows the structure of the experiment that includes blocking using age is shown in Figure 2.9.

Figure 2.9
Diagram for the experiment of Example 2.9 using age to form blocks.


5049 participants
Create blocks
789 age 18–29
1532 age 30–44
1089 age 45–59
1639 age 60+
Random assignment
No labels
Low-impact labels
High-impact labels
Record food choice
Compare percent ordering red meat for this age group
Compare percent ordering red meat for the three menu types

When blocking is used, the design is called a **randomized block design**. Notice that one difference between the diagram that describes the experiment in which blocking is used (Figure 2.9) and the diagram of the original experiment (Figure 2.5) is the point where the random assignment occurs. When blocking is incorporated in an experiment, the random assignment to treatments occurs *after* the blocks have been formed and is done separately for each block.

Before proceeding with an experiment, you should be able to give a satisfactory answer to each of the following 10 questions.

1. What is the research question that data from the experiment will be used to answer?
2. What is the response variable?

Copyright 2025 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s). Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

3. How will the values of the response variable be determined?
4. What are the explanatory variables for the experiment?
5. For each explanatory variable, how many different values are there, and what are these values?
6. What are the treatments for the experiment?
7. What extraneous variables might influence the response?
8. How does the design incorporate random assignment of subjects to treatments (or treatments to subjects) or random assignment of treatments to trials?
9. For each extraneous variable listed in Question 7, does the design protect against its potential influence on the response through blocking, direct control, or random assignment?
10. Will you be able to answer the research question using the data collected in this experiment?

Exercises 2.35–2.51

2.35 The head of the quality control department at a printing company would like to carry out an experiment to determine which of three different glues results in the greatest binding strength. Although they are not of interest in the current study, other factors thought to affect binding strength are the number of pages in the book and whether the book is being bound as a paperback or a hardback. (Hint: See box on page 47.)

- **a.** What is the response variable in this experiment?
- **b.** What explanatory variable will determine the experimental conditions?
- **c.** What two extraneous variables are mentioned in the problem description? Can you think of any other extraneous variables that should be considered?

2.36 A study of college students showed a temporary gain of up to 9 IQ points after listening to a Mozart piano sonata. This conclusion, dubbed the Mozart effect, has since been criticized by a number of researchers who have been unable to confirm the result in similar studies. Suppose that you wanted to see whether there is a Mozart effect for students at your school. (Hint: See Examples 2.4 and 2.5.)

- **a.** Describe how you might design an experiment for this purpose.
- **b.** Does your experimental design include direct control of any extraneous variables? Explain.
- **c.** Does your experimental design use blocking? Explain why you did or did not include blocking in your design.
- **d.** What role does random assignment play in your design?

2.37 According to the article **"Rubbing Hands Together Under Warm Air Dryers Can Counteract Bacteria Reduction" (*Infectious Disease News*, September 22, 2010)** washing your hands isn't enough—good "hand hygiene" also includes drying hands thoroughly. The article described an experiment to compare bacteria reduction for three different hand-drying methods. In this experiment, subjects handled uncooked chicken for 45 seconds, then washed their hands with a single squirt of soap for 60 seconds, and then used one of the three hand-drying methods. After completely drying their hands, the bacteria count on their hands was measured.

Suppose you want to carry out a similar experiment and that you have 30 subjects who are willing to participate. Describe a method for randomly assigning each of the 30 subjects to one of the hand-drying methods. (Hint: See A Note on Random Assignment on page 51.)

2.38 Which is a better approach for taking notes in class: a laptop or a pen and paper? The author of the article **"Laptops Are Great. But Not During a Lecture or a Meeting" (*New York Times*, November 22, 2017)** argues that college students learn less and earn worse grades when they use computers or tablets during lecture. They based this conclusion on several studies, one of which was an experiment in which 67 students at Princeton University were asked to take notes while watching videos about topics that were "interesting but not common knowledge." Students were randomly assigned to a note-taking method: using a laptop or using a notebook. Thirty minutes later, all participants took the same test to measure their knowledge of the material covered in the videos. On factual-recall questions, there was little difference between the two groups, but on conceptual-application questions, laptop users performed significantly worse. Explain why it was important

Copyright 2025 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s). Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

for the researchers to randomly assign the subjects to the two groups.

2.39 The report **"Comparative Study of Two Computer Mouse Designs" (Cornell Human Factors Laboratory Technical Report RP7992)** describes an experiment in which 24 volunteers were divided into three groups based on height. The volunteers in each group were then randomly assigned to use one of two different computer mouse designs. The study only included people who reported that they used their right hand to operate a computer mouse.

This experimental design incorporated direct control and blocking.

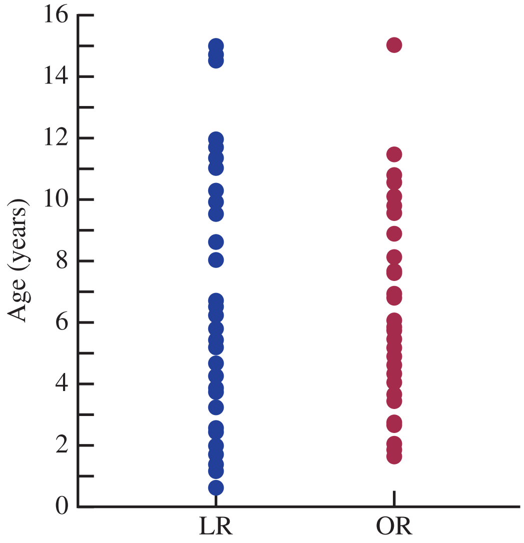
a. Is the potential effect of the extraneous variable stature (height) addressed by blocking or direct control?

b. Whether the right or left hand is used to operate the mouse was considered to be an extraneous variable. Is the potential effect of this variable addressed by blocking or direct control?

2.40 An article titled **"The Discipline of Business Experimentation" (*Harvard Business Review*, December 2014)** provides several examples of randomized experiments in the context of making business decisions. In one such experiment, the retailer Kohl's decided to test whether opening stores an hour later would lead to a significant decrease in sales. The experiment involved 100 of the company's stores, which were divided into two groups. One group of stores opened an hour later than usual. The other group continued with regular store hours. The experiment did not find a meaningful difference in sales between the groups of stores. Explain why it is important that the company use random assignment to create the two experimental groups.

2.41 In an experiment to compare two different surgical procedures for hernia repair **("A Single-Blinded, Randomized Comparison of Laparoscopic Versus Open Hernia Repair in Children," *Pediatrics* [2009]: 332–336)**, 89 children were assigned at random to one of the two surgical methods. The researchers relied on the random assignment of subjects to treatments to create comparable groups with respect to extraneous variables that they did not control. One such extraneous variable was age.

After random assignment to treatments, the researchers looked at the age distribution of the children in each of the two experimental groups (laparoscopic repair [LR] and open repair [OR]). The accompanying figure is similar to one in the paper.



Based on this figure, has the random assignment of subjects to experimental groups been successful in creating groups that are similar with respect to the ages of the children in the groups? Explain.

2.42 The article **"Vouchers Found to Lower Test Scores in Washington Schools" (*New York Times*, April 28, 2017)** describes and experiment to investigate the effect of providing tuition vouchers to low-income parents who want their child to attend a private school. A total of 1700 students applied to a school voucher program. These students were assigned at random to a group that received the vouchers and a group that did not.

a. The researchers concluded that the students in the group that received a voucher scored significantly lower in mathematics achievement than students in the group that did not receive a voucher. Explain why the researchers would not have been able to attribute the lower scores to receiving a voucher if all of the students applying to the program were given vouchers and then this group was compared to students who did not receive vouchers because they did not apply to the program.

b. Construct a diagram to represent the underlying structure of this experiment.

2.43 Many surgeons play music in the operating room. Does the type of music played have an effect on the surgeons' performance? The report **"Death Metal in the Operating Room" (npr.org, December 24, 2009)** describes an experiment in which surgeons used a simulator to perform a surgery. Some of the surgeons listened to music with vocal elements while performing the surgery, and others listened to music that did not have vocal elements. The researchers concluded that the average time to complete the surgery was greater when music with

Copyright 2025 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s). Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

vocal elements is played than when music without vocal elements is played.

a. What are the experimental conditions for the experiment described above? What is the response variable?

b. Explain why it was important to control each of the following variables in the experiment described above.

  i. Type of surgery performed

  ii. Operating room temperature

  iii. Volume at which the music was played

c. Explain why it is important that the surgeons be assigned at random to the two music conditions in the experiment described above.

2.44 Construct a diagram to represent the note-taking experiment of Example 2.4.

2.45 Construct a diagram to represent the gasoline additive experiment described on page 49.

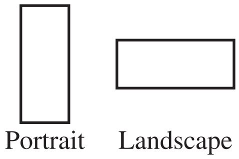
2.46 An advertisement for a sweatshirt that appeared in ***SkyMall Magazine*** (a catalog distributed by some airlines) stated the following: "This is not your ordinary hoody! Why? Fact: Research shows that written words on containers of water can influence the water's structure for better or worse depending on the nature and intent of the word. Fact: The human body is 70% water. What if positive words were printed on the inside of your clothing?"

The reference to the "fact" that written words on containers of water can influence the water's structure appears to be based on the work of Dr. Masaru Emoto who typed words on paper, pasted the words on bottles of water, and observed how the water reacted to the words by seeing what kind of crystals were formed in the water. Dr. Masaru describes several of his experiments in the self-published book, ***The Message from Water.***

If you were going to interview Dr. Emoto, what questions would you want to ask about the design of this experiment?

2.47 The paper **"Turning to Learn: Screen Orientation and Reasoning from Small Devices" (*Computers in Human Behavior* [2011]: 793–797)** describes a study that investigated whether cell phones with small screens are useful for gathering information. The researchers wondered if the ability to reason using information read on a small screen was affected by the screen orientation. The researchers assigned 33 undergraduate students who were enrolled in a psychology course at a large public university to one of two groups at random.

One group read material that was displayed on a small screen in portrait orientation, and the other group read material on the same size screen but turned to display the information in landscape orientation (see the following figure).



The researchers found that performance on a reasoning test that was based on the displayed material was better for the group that read material in the landscape orientation.

a. Is the described study an observational study or an experiment?

b. Did the study use random selection from some population?

c. Did the study use random assignment to experimental groups?

2.48 Consider the study described in the previous exercise.

a. Is the conclusion—that reasoning using information displayed on a small screen is improved by turning the screen to landscape orientation—appropriate, given the study design? Explain.

b. Is it reasonable to generalize the conclusions from this study to some larger population? If so, what population?

2.49 Does wearing glasses decrease the risk of infection with respiratory viruses? The paper **"Effect of Wearing Glasses on Risk of Infection With SARS-CoV-2 in the Community" (*JAMA Network Open*, [2022]: e2244495)** describes an experiment in which 3717 adults who did not regularly wear glasses, had no symptoms of COVID-19, and who had not had COVID-19 in the previous 2 weeks were randomly assigned to one of two groups. One group wore glasses when in public spaces for a period of 2 weeks and those in the other group did not wear glasses. At the end of the 2-week period, the percentages who had tested positive during the time period for the two groups were compared.

a. What are the two treatments in this experiment?

b. What is the response variable in this experiment?

2.50 A study of college students showed a temporary gain of up to nine IQ points after listening to a Mozart piano sonata. This result, dubbed the Mozart effect, has since been criticized by a number of researchers who have been unable to confirm the result in similar studies. Suppose that you want to determine if there really is a Mozart effect. You decide to carry out an experiment with three experimental groups. One group will listen to a Mozart piano sonata that lasts for 24-minutes. The second group will listen to popular music for the same length of time, and the third group will relax for 24-minutes with no music playing. You will measure IQ before and after the 24-minute period. Suppose that you have 45 volunteers who have agreed

Copyright 2025 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s). Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

to participate in the experiment. Describe the steps in a process you could use to randomly assign each of the volunteers to one of the experimental groups.

2.51 Researchers at the University of Utah carried out a study to see if the size of the fork used to eat dinner has an effect on how much food is consumed (***Food Network Magazine*, January 2012**). The researchers assigned people to one of two groups. One group ate dinner using a small fork, and the other group ate using a large fork. The researchers found that those who ate with a large fork ate less of the food on the plate than those who ate with the small fork. The title of the article describing this study was **"Dieters Should Use a Big Fork."** This title implies a cause-and-effect relationship between fork size and amount eaten and also generalizes this finding to the population of dieters. What would you need to know about the study design to determine if the conclusions implied by the headline are reasonable?

Section 2.4 More on Experimental Design

The previous section covered basic principles for designing simple comparative experiments—control, blocking, random assignment, and replication. The goal of an experimental design is to provide a method of data collection that

1. minimizes the effect of extraneous sources of variability in the response so that any differences in response for various experimental conditions can be more easily assessed and
2. creates experimental groups that are similar with respect to extraneous variables that cannot be controlled either directly or through blocking.

In this section, some additional things to think about when planning an experiment are considered.

Use of a Control Group

If the purpose of an experiment is to determine whether some treatment has an effect, it is important to include an experimental group that does not receive the treatment. Such a group is called a **control group**. The use of a control group allows the experimenter to assess how the response variable behaves when the treatment is not used. This provides a baseline against which the treatment groups can be compared to determine whether the treatment had an effect.

Example 2.10 Comparing Gasoline Additives

Suppose that an engineer wants to know whether a gasoline additive increases fuel efficiency (miles per gallon). The experiment might use a single car (to eliminate car-to-car variability) and a sequence of trials in which 1 gallon of gas is put in an empty tank, the car is driven around a racetrack at a constant speed, and the distance traveled on the gallon of gas is recorded.

What Do You Think?

Would you be able to determine if the additive increases fuel efficiency if there were not trials where the additive was not used?

To determine whether the additive increases gas mileage, it would be necessary to include a control group of trials in which distance traveled was measured when gasoline without the additive was used. The trials would be assigned *at random* to one of the two experimental conditions (additive or no additive [the control group]).

Even though this experiment consists of a sequence of trials all with the same car, random assignment of trials to experimental conditions is still important because there might be extraneous variables that could affect fuel efficiency. For example, temperature or other environmental conditions might change over the sequence of trials, or the physical condition of the car might change slightly from one trial to another. Random assignment of experimental conditions to trials will tend to even out the effects of these extraneous variables.

Copyright 2025 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s). Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

Although a control group is usually one that receives no treatment, in experiments designed to compare a new treatment to an existing standard treatment, the term control group is sometimes also used to describe the group that receives the current standard treatment.

Not all experiments require the use of a control group. Many experiments are designed to compare two or more conditions—for example, an experiment to determine how oven temperature affects the cooking time of a particular type of cake. However, sometimes a control group is included even when the ultimate goal is to compare two or more different treatments. This is because an experiment with two treatments and no control group might allow you to determine whether there is a difference between the two treatments and even to assess the magnitude of the difference if one exists, but it would not allow you to assess the individual effect of either treatment. For example, without a control group, you might be able to say that there is no difference in the increase in mileage for two different gasoline additives, but you would not be able to tell if this was because both additives increased gas mileage by a similar amount or because neither additive had any effect on gas mileage.

Use of a Placebo

In experiments that use human subjects, use of a control group may not be enough to determine whether a treatment really does have an effect. People sometimes respond merely to the power of suggestion! For example, suppose a study designed to determine whether a particular herbal supplement is effective in promoting weight loss uses an experimental group that takes the herbal supplement and a control group that takes nothing. It is possible that those who take the herbal supplement and believe that they are taking something that will help them to lose weight may be more motivated and may unconsciously change their eating behavior or activity level, resulting in weight loss.

Although there is debate about the degree to which people respond, many studies have shown that people sometimes respond to treatments with no active ingredients and that they often report that such “treatments” relieve pain or reduce symptoms. So, if researchers want to determine whether a treatment really has an effect, comparing a treatment group to a control group may not be enough. To address this problem, many experiments use what is called a **placebo**.

Definition

Placebo: Something that is identical (in appearance, taste, feel, etc.) to the treatment received by the treatment group, except that it contains no active ingredients.

What Do You Think?

For the herbal supplement experiment, would it be helpful to include a group that received a placebo?

For example, in the herbal supplement experiment, rather than using a control group that received *no* treatment, the researchers might want to include a placebo group. Individuals in the placebo group would take a pill that looked just like the herbal supplement but did not contain the herb or any other active ingredient. As long as the subjects did not know whether they were taking the herb or the placebo, the placebo group would provide a better basis for comparison and would allow the researchers to determine whether the herbal supplement had any real effect over and above the “placebo effect.”

Single-Blind and Double-Blind Experiments

Because people often have their own personal beliefs about the effectiveness of various treatments, it is sometimes desirable to conduct experiments in a way that subjects do not know what treatment they are receiving. For example, in an experiment comparing four different doses of a medication for relief of headache pain, someone who knows that they are receiving the medication at its highest dose may be subconsciously influenced to report a greater degree of headache pain reduction. By ensuring that subjects are not aware of which treatment they receive, you can prevent the subjects' personal perceptions from influencing the response.

An experiment in which subjects do not know what treatment they have received is described as **single-blind**. Of course, not all experiments can be made single-blind. For example, in an experiment to compare the effect of two different types of exercise on blood pressure, it is not possible for participants to be unaware of whether they are in the swimming group or the jogging group! However, when it is possible, "blinding" the subjects in an experiment is generally a good strategy.

In some experiments, someone other than the subject is responsible for measuring the response. To ensure that the person measuring the response does not let personal beliefs influence the way in which the response is recorded, the researchers should make sure that the measurer does not know which treatment was given to any particular individual. For example, in a medical experiment to determine whether a new vaccine reduces the risk of getting the flu, doctors must decide whether a particular individual who is not feeling well actually has the flu or some other unrelated illness. If the doctor knew that a participant with flu-like symptoms had received the new flu vaccine, they might be less likely to determine that the participant had the flu and more likely to interpret the symptoms as being the result of some other illness.

There are two ways in which blinding might occur in an experiment. One involves blinding the subjects, and the other involves blinding the individuals who measure the response. If subjects do not know which treatment was received *and* those measuring the response do not know which treatment was given to which subject, the experiment is described as **double-blind**. If only one of the two types of blinding is present, the experiment is **single-blind**.

Definitions

Double-blind experiment: An experiment in which neither the subjects nor the individuals who measure the response know which treatment was received.

Single-blind experiment: An experiment in which the subjects do not know which treatment was received but the individuals measuring the response do know which treatment was received, or one in which the subjects do know which treatment was received but the individuals measuring the response do not know which treatment was received.

Experimental Units and Replication

An **experimental unit** is the smallest unit to which a treatment is applied. In the language of experimental design, treatments are assigned at random to experimental units, and replication means that each treatment is applied to more than one experimental unit.

Replication is necessary for random assignment to be an effective way to create similar experimental groups and to get a sense of the variability in the values of the response for individuals who receive the same treatment. As you will see in Chapters 9–15, this enables you to use statistical methods to decide whether differences in the responses in different treatment groups can be attributed to the treatment received or whether they can be explained by chance variation (the natural variability seen in the responses to a single treatment).

Be careful when designing an experiment to ensure that there is replication. For example, suppose that children in two third-grade classes are available to participate in an experiment to compare two different methods for teaching arithmetic. It might at first seem reasonable to select one class at random to use one method and then assign the other method to the remaining class. But what are the experimental units here? If treatments are randomly assigned to classes, classes are the experimental units. Because only one class is assigned to each treatment, this is an experiment with no replication, even though there are many children in each class. You would *not* be able to determine whether there was a difference between the two methods based on data from this experiment, because you would have only one observation per treatment.

One last note on replication: Do not confuse replication in an experimental design with replicating an experiment. Replicating an experiment means conducting a new experiment using the same experimental design as a previous experiment. This is a way of confirming conclusions based on a previous experiment, but it does not eliminate the need for replication in each of the individual experiments themselves.

Using Volunteers as Subjects in an Experiment

Although the use of volunteers in a study that involves collecting data through sampling is never a good idea, it is a common practice to use volunteers as subjects in an experiment. Even though the use of volunteers limits the researcher's ability to generalize to a larger population, random assignment of the volunteers to treatments should result in comparable groups, and so treatment effects can still be assessed.

Exercises 2.52–2.63

2.52 Explain why some studies include both a control group and a placebo treatment. What additional comparisons are possible if both a control group and a placebo group are included?

2.53 Explain why blinding is a reasonable strategy in many experiments.

2.54 Give an example of an experiment for each of the following:

a. Single-blind experiment with the subjects blinded

b. Single-blind experiment with the individuals measuring the response blinded

c. Double-blind experiment

d. An experiment for which it is not possible to blind the subjects

2.55 The article **"Study Points to Benefits of Knee Replacement Surgery Over Therapy Alone" (*The New York Times*, October 21, 2015)** describes a study to compare two treatments for people with knee pain. In the study, 50 people with arthritis received knee replacement surgery followed by a program of exercise. Another 50 people with arthritis did not have surgery but received the same program of exercise. After 1 year, 85% of the people who had surgery and 68% of the people who did not have surgery reported pain relief.

a. Why would it be important to determine if the researchers randomly assigned the people participating in the study to one of the two groups?

b. Explain why you think that the researchers did not include a control group (a group that did not receive surgery and also did not receive an exercise program) in this study.

2.56 Which is a better approach for taking notes in class: a laptop or a pen and paper? The author of the article **"Laptops Are Great. But Not During a Lecture or a Meeting" (*New York Times*, November 22, 2017)** argues that college students learn less and earn worse grades when they use computers or tablets during lecture. They based this conclusion on several studies, one of which was an experiment in which 67 students at Princeton University were asked to take notes while watching videos about topics that were "interesting but not common knowledge." Students were randomly assigned to a note-taking method: using a laptop or using a notebook. Thirty minutes later, all participants took the same test to measure their knowledge of the material covered in the videos. On factual-recall questions, there was little difference between the two groups, but on conceptual application questions, laptop users performed significantly worse. The article does not mention blinding, but the full report, published in *Psychological Science* (2014), includes this description: "An independent rater, blind to the purpose of the study and condition, scored all open-ended questions."

a. Why do you think it was important that the person who scored the open-ended questions did not know whether the student took notes on a laptop or in a notebook?

b. Is it possible for this experiment to be double-blind? Explain why or why not.

2.57 The article **"Placebos Are Getting More Effective. Drug Makers Are Desperate to Know Why." (*Wired Magazine*, August 8, 2009)** states that "according to research, the color of a tablet can boost the effectiveness even of genuine meds—or help convince a patient that a placebo is a potent remedy." Describe how you would design an experiment to investigate if adding color to Tylenol tablets would result in greater perceived pain relief. Be sure to address how you would select subjects, how you would measure pain relief, what colors you would use, and whether or not you would include a control group in your experiment.

2.58 The article **"Yes That Miley Cyrus Biography Helps Learning" (*The Globe and Mail,* August 5, 2010)** describes an experiment investigating whether providing summer reading books to children would affect school performance. Subjects in the experiment were 1330 children randomly selected from first and second graders at selected schools in Florida. A group of 852 of these children were selected at random from the group of 1330 participants to be in the "book" group. The other 478 children were assigned to the control group.

Children in the book group were invited to a book fair in the spring to choose any 12 reading books which they could then take home. Children in the control group were not given any reading books but were given some activity and puzzle books. This process was repeated each year for 3 years until the children reached third and fourth grade. The researchers then compared reading test scores of the two groups.

a. Do you think that randomly selecting 852 of the 1330 children to be in the book group is equivalent to random assignment of the children to the two experimental groups? Explain.

b. Explain the purpose of including a control group in this experiment.

2.59 Suppose that the researchers who carried out the experiment described in the previous exercise thought that grade level at the start of the study (first or second grade) might be a potentially confounding variable. If 700 of the children participating in the experiment were in first grade and 630 were in second grade, describe how blocking could be incorporated into the experiment. Be specific about how you would assign the children to treatment groups.

2.60 The article **"The Science of Sniff: Why Dogs Are Great Disease Detectors" (*Wired*, October 30, 2018)** refers to an experiment in which dogs were trained to distinguish between children with malaria and healthy children by sniffing clothes worn by the children. Dogs were trained to walk down a line of children, pausing to sniff each one and sitting down if they detected the scent of malaria. After training, a lab test found that the canines could correctly identify children infected with malaria 70% of the time. The article specified that the test was blinded, meaning that the dog handlers and those observing the experiment did not know which children were infected with malaria. Explain why blinding is an important aspect of the design of this experiment.

2.61 Pismo Beach, California, has an annual clam festival that includes a clam chowder contest. Judges rate clam chowders from local restaurants, and the judging is done in such a way that the judges are not aware of which chowder is from which restaurant. One year, much to the dismay of the seafood restaurants on the waterfront, Denny's chowder was declared the winner! (When asked what the ingredients were, the cook at Denny's said they weren't sure—they just had to add the right amount of nondairy creamer to the soup stock that they got from Denny's distribution center!)

a. Do you think that Denny's chowder would have won the contest if the judging had not been "blind"? Explain.

b. Although this was not an experiment, your answer to Part (a) helps to explain why those measuring the response in an experiment are often blinded. Using your answer in Part (a), explain why experiments are often blinded in this way.

2.62 The paper **"Effect of a Nutritional Supplement on Hair Loss in Women" (*Journal of Cosmetic Dermatology* [2015]: 76–82)** describes an experiment to see if a dietary supplement consisting of Omega-3, Omega-6, and antioxidants could reduce hair loss in women with stage 1 hair loss. One hundred twenty women volunteered to participate in the study and were randomly assigned to either the supplement group or a control group. The women in the supplement group took the supplement for 6 months. Photos of the top of the head were taken of all the women at the beginning of the study and 6 months later at the end of the study. The two photos of each woman were evaluated by an independent expert who visually determined the change in hair density. The expert who determined the change in hair density did not know which of the women had taken the supplement.

Answer the following seven questions for this experiment.

1. What question is the experiment trying to answer?
2. What are the experimental conditions (treatments) for this experiment?
3. What is the response variable?
4. What are the experimental units and how were they selected?
5. Does the design incorporate random assignment of experimental units to the different experimental conditions? If not, are there potentially confounding variables that would make it difficult to draw conclusions based on data from the experiment?
6. Does the experiment incorporate a control group and/or a placebo group? If not, would the experiment be improved by including one or both of these?

Copyright 2025 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s). Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

7. Does the experiment involve blinding? If not, would the experiment be improved by making it single- or double-blind?

2.63 The article **"A Debate in the Dentist's Chair" (*San Luis Obispo Tribune,* January 28, 2000)** described an ongoing debate over whether newer resin fillings are a better alternative to the more traditional silver amalgam fillings. Because amalgam fillings contain mercury, there is concern that they could be mildly toxic and prove to be a health risk to those with some types of immune and kidney disorders. One experiment described in the article used sheep as subjects and reported that sheep treated with amalgam fillings had impaired kidney function.

a. In the experiment, a control group of sheep that received no fillings was used but there was no placebo group. Explain why it is not necessary to have a placebo group in this experiment.

b. The experiment compared only an amalgam filling treatment group to a control group. What would be the benefit of also including a resin filling treatment group in the experiment?

c. Why do you think the experimenters used sheep rather than human subjects?

Section 2.5 Interpreting and Communicating the Results of Statistical Analyses

Statistical studies are conducted to answer questions about characteristics of some population of interest or about the effect of some treatment. These questions are answered using data, and how the data are obtained determines the quality of information available and the type of conclusions that can be drawn. As a consequence, when describing a study you have conducted (or when evaluating a published study), it is important to consider how the data were collected.

The description of the data collection process should make it clear whether the study is an observational study or an experiment. For observational studies, some of the issues that should be addressed are:

1. What is the population of interest? What is the sampled population? Are these two populations the same? If the sampled population is only a subset of the population of interest, **undercoverage** limits the ability to generalize to the population of interest. For example, if the population of interest is all students at a particular university, but the sample is selected from only those students who choose to list their phone number in the campus directory, undercoverage may be a problem. You would need to think carefully about whether it is reasonable to consider the sample as representative of the population of all students at the university.

   Overcoverage results when the sampled population is actually larger than the population of interest. This would be the case if you were interested in the population of all high schools that offer Advanced Placement (AP) Statistics but sampled from a list of all schools that offered an AP class in any subject. Both undercoverage and overcoverage can be problematic.
2. How were the individuals or objects in the sample actually selected? A description of the sampling method helps the reader to make judgments about whether the sample can reasonably be viewed as representative of the population of interest.
3. What are potential sources of bias, and is it likely that any of these will have a substantial effect on the observed results? When describing an observational study, you should acknowledge that you are aware of potential sources of bias and explain any steps that were taken to minimize their effect. For example, in a mail survey, nonresponse can be a problem, but the sampling plan may seek to minimize its effect by offering incentives for participation and by following up one or more times with those who do not respond to the first request.

   A common misperception is that increasing the sample size is a way to reduce bias in observational studies, but this is not the case. For example, if measurement bias is present, as in the case of a scale that is not correctly calibrated and tends to weigh too high, taking 1000 measurements rather than 100 measurements cannot correct for the fact that the measured weights will be too large. Similarly, a larger sample size cannot compensate for response bias introduced by a poorly worded question.

Copyright 2025 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s). Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

For experiments, some of the issues that should be addressed are:

1. What is the role of random assignment? All good experiments use random assignment as a means of coping with the effects of potentially confounding variables that cannot easily be directly controlled. When describing an experimental design, you should be clear about how random assignment (subjects to treatments, treatments to subjects, or treatments to trials) was incorporated into the design.
2. Were any extraneous variables directly controlled by holding them at fixed values throughout the experiment? If so, which ones and at which values?
3. Was blocking used? If so, how were the blocks created? If an experiment uses blocking to create groups of similar experimental units, the criteria used to create the blocks should be described and an explanation of why these blocks were used should be given. For example, you might say something like "Subjects were divided into two blocks—those who exercise regularly and those who do not exercise regularly—because it was believed that exercise status might affect the responses to the diets."

Because each treatment appears at least once in each block, the block size must be at least as large as the number of treatments. Ideally, the block sizes should be equal to the number of treatments, because this presumably would allow the experimenter to create small groups of very similar experimental units. For example, in an experiment to compare two methods for teaching calculus to first-year college students, you may want to block on previous mathematics knowledge by using math SAT scores. If 100 students are available as subjects for this experiment, rather than creating two large groups (above-average math SAT score and below-average math SAT score), you might want to create 50 blocks of two students each, the first consisting of the two students with the highest math SAT scores, the second containing the two students with the next highest scores, and so on. You would then select one student in each block at random and assign that student to teaching method 1. The other student in the block would be assigned to teaching method 2.

A Word to the Wise: Cautions and Limitations

It is a mistake to begin collecting data before thinking carefully about research objectives and developing a plan. A poorly designed plan for data collection may result in data that do not enable the researcher to answer key questions of interest or to generalize conclusions based on the data to the desired populations of interest.

Clearly defining the objectives at the outset enables the investigator to determine whether an experiment or an observational study is the best way to proceed. Watch out for the following *inappropriate* actions:

1. Drawing a cause-and-effect conclusion from an observational study. Don't do this, and don't believe it when others do it!
2. Generalizing results of an experiment that uses volunteers as subjects to a larger population. This is not sensible without a convincing argument that the group of volunteers can reasonably be considered to be representative of the population.
3. Generalizing conclusions based on data from a sample to some population of interest. This is sometimes a sensible thing to do, but on other occasions it is not reasonable. Generalizing from a sample to a population is justified only when there is reason to believe that the sample is likely to be representative of the population. This would be the case if the sample was a random sample from the population and there were no major potential sources of bias. If the sample was not selected at random or if potential sources of bias were present, these issues would have to be addressed before a judgment could be made regarding the appropriateness of generalizing the study results.

   For example, the media company Stacker gave a list of the median 2022 housing price for each of the 30 cities in California with the highest home prices **("Cities With the Most Expensive Homes in California," stacker.com, 2022)**. Although these 30 cities are a sample of the cities in California, they were not randomly selected and (because they are the 30 cities with the highest house prices) it would not be reasonable to generalize to all California cities bases on data from this sample.
4. Generalizing conclusions based on an observational study that used voluntary response or convenience sampling to a larger population. This is almost never reasonable.

Copyright 2025 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s). Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

Exercises 2.64–2.68

2.64 The following appeared in a ***Psychology Today*** blog post **(December 27, 2022):**

> Every December, the *British Medical Journal (BMJ)* editors put together a light-hearted Christmas issue that contains playful evidence-based scientific studies that are rigorously researched and peer-reviewed—but also meant to be fun.
>
> This year, *BMJ*'s Christmas issue features a comedy-inspired study (Gaesser et al., 2022) on how walking in a highly inefficient style—à la Monty Python's Mr. Teabag from the "Ministry of Silly Walks" TV skit—turns people into calorie-burning machines.
>
> The researchers found that just 11 minutes per day (77 minutes/week) of silly walking—which involves randomly throwing one leg up in the air, followed by spontaneous deep knee bends or little bunny hops—fulfills the cardio portion of someone's weekly exercise needs based on the most-recent physical activity guidelines.

The actual study referred to in this blog post is described in the paper **"Quantifying the Benefits of Inefficient Walking: Monty Python Inspired Laboratory Based Experimental Study" (*British Medical Journal* [2022]: e072833)**. You can access this paper at www.bmj.com/content/379/bmj-2022-072833. If you aren't familiar with the Monty Python Ministry of Silly Walks skit, you can watch it at www.youtube.com/watch?v=TNeeovY4qNU.

After watching the video and reading the paper, write up a short summary of the experiment and the conclusions drawn from it. Include a description of any important aspects of the experiment and whether you think that the experiment was conducted in a way that supports the conclusions given.

2.65 The article **"Effects of Too Much TV Can Be Undone" (*USA TODAY*, October 1, 2007)** included the following paragraph:

> Researchers at Johns Hopkins Bloomberg School of Public Health report that it's not only how many hours children spend in front of the TV, but at what age they watch that matters. They analyzed data from a national survey in which parents of 2707 children were interviewed first when the children were 30–33 months old and again when they were $5\frac{1}{2}$ years old, about their TV viewing and their behavior.

a. Is the study described an observational study or an experiment?

b. The article says that data from a sample of 2707 parents were used in the study. What other information about the sample would you want in order to evaluate the study?

2.66 The actual paper referred to in the ***USA TODAY*** article described in the previous exercise was **"Children's Television Exposure and Behavioral and Social Outcomes at 5.5 years: Does Timing of Exposure Matter?" (*Pediatrics* [2007]: 762–769).** The paper describes the sample as follows:

> The study sample included 2707 children whose mothers completed telephone interviews at both 30 to 33 months and 5.5 years and reported television exposure at both time points. Of those completing both interviewers, 41 children (1%) were excluded because of missing data on television exposure at one or both time points. Compared with those enrolled in the HS clinical trial, parents in the study sample were disproportionately older, white, more educated, and married.

a. The "HS clinical trial" referred to in the excerpt from the paper was a nationally representative sample used in the Healthy Steps for Young Children national evaluation. Based on the above description of the study sample, do you think that it is reasonable to regard the sample as representative of parents of all children at age 5.5 years? Explain.

b. The ***USA TODAY*** article also includes the following summary paragraph:

> The study did not examine what the children watched and can't show TV was the cause of later problems, but it does "tell parents that even if kids are watching TV early in life, and they stop, it could reduce the risk for behavioral and social problems later," Mistry says.

What potentially confounding variable is identified in this passage?

c. The passage in Part (b) says that the study cannot show that TV was the cause of later problems. Is the quote from Kamila Mistry (one of the study authors) in the passage consistent with the statement about cause? Explain.

2.67 An article titled **"I Said, Not While You Study: Science Suggests Kids Can't Study and Groove at the Same Time"** appeared in the ***Washington Post*** **(September 5, 2006)**. This provides an example of a reporter summarizing the result of a scientific study in a way that is designed to make it accessible to the newspaper's readers. You can find the newspaper article online by searching on the title or by going to **washingtonpost.com/wp-dyn/content/article/2006/09/03/AR2006090300592.html**. The study referenced in the newspaper article was published in the ***Proceedings of the National Academies of Science*** and can be found at **pnas.org/content/103/31/11778.full**.

Read the newspaper article and then take a look at the published paper. Comment on whether you think that the author was successful in communicating the findings of the study to the intended audience.

2.68 The short article **"Developing Science-Based Food and Nutrition Information" (*Journal of the American Dietetic Association* [2001]: 1144–1145)** includes some guidelines for evaluating a research paper. Obtain a copy of this paper through your university library or your instructor. Read this article and make a list of questions that can be used to evaluate a research study.

Chapter Activities

Activity 2.1 Facebook Friending

The purpose of this activity is to provide an opportunity to design a sampling plan.

Background: The article **"Professors Prefer Face Time to Facebook"** appeared in the student newspaper at Cal Poly, San Luis Obispo **(*Mustang Daily*, August 27, 2009)**. The article examines how professors and students felt about using Facebook as a means of faculty-student communication. The student who wrote this article got mixed opinions when they interviewed students to ask whether they wanted to become Facebook friends with their professors. Two student comments included in the article were

> "I think the younger the professor is, the more you can relate to them and the less awkward it would be if you were to become friends on Facebook. The older the professor, you just would have to wonder, 'Why are they friending me?' "

and

> "I think becoming friends with professors on Facebook is really awkward. I don't want them being able to see into my personal life, and frankly, I am not really interested in what my professors do in their free time."

Even if the students interviewed had expressed a consistent opinion, it would still be unreasonable to think this represented general student opinion on this issue because only four students were interviewed and it is not clear from the article how these students were selected.

In this activity, you will work with a partner to develop a plan to assess student opinion about connecting with professors over social media at your school.

1. Suppose you will select a sample of 50 students at your school to participate in a survey. Write one or more questions that you would ask each student in the sample.
2. Discuss with your partner whether you think it would be easy or difficult to obtain a simple random sample of 50 students at your school and to obtain the desired information from all the students selected for the sample. Write a summary of your discussion.
3. With your partner, decide how you might go about selecting a sample of 50 students from your school that reasonably could be considered representative of the population of interest even if it may not be a simple random sample. Write a brief description of your sampling plan, and point out the aspects of your plan that you think make it reasonable to argue that it will be representative.
4. Explain your plan to another pair of students. Ask them to critique your plan. Write a brief summary of the comments you received. Now reverse roles, and provide a critique of the plan devised by the other pair.
5. Based on the feedback you received in Step 4, would you modify your original sampling plan? If not, explain why this is not necessary. If so, describe how the plan would be modified.

Activity 2.2 An Experiment to Test for the Stroop Effect

The purpose of this activity is to provide an opportunity to design and carry out an experiment.

Background: In 1935, John Stroop published the results of his research into how people respond when presented with conflicting signals. Stroop noted that most people are able to read words quickly and that they cannot easily ignore them and focus on other attributes of a printed word, such as text color. For example, consider the following list of words:

green blue red blue yellow red

It is easy to quickly read this list of words. It is also easy to read the words even if the words are printed in color, and even if the text color is different from the color of the word. For example, people can read the words in a list of words that are printed in different colors like the list shown here

green blue red blue yellow red

as quickly as they can read the list that isn't printed in color.

However, Stroop found that if people are asked to name the text colors of the words in the list (red, yellow, blue, green, red, green), it takes them longer. Psychologists believe that this is because the reader has to inhibit a

Copyright 2025 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s). Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

natural response (reading the word) and produce a different response (naming the color of the text).

If Stroop is correct, people should be able to name colors more quickly if they do not have to inhibit the word response, as would be the case if they were shown colored rectangles like those shown below.

1. Design an experiment to compare times to identify colors when they appear as text to times to identify colors when there is no need to inhibit a word response. Indicate how random assignment is incorporated into your design. What is your response variable? How will you measure it? How many subjects will you use in your experiment, and how will they be chosen?
2. When you are satisfied with your experimental design, carry out the experiment. You will need to construct a list of colored words and a corresponding list of colored bars to use in the experiment. You will also need to think about how you will implement the random assignment scheme.
3. Summarize the resulting data in a brief report that explains whether your findings are consistent with the Stroop effect.

Activity 2.3 McDonald's and the Next 100 Billion Burgers

The purpose of this activity is to provide experience in evaluating data collection methods.

Background: The article **"Potential Effects of the Next 100 Billion Hamburgers Sold by McDonald's" (*American Journal of Preventative Medicine* [2005]: 379–381)** estimated that 992.25 million pounds of saturated fat would be consumed as McDonald's sells its next 100 billion hamburgers. This estimate was based on the assumption that the average weight of a burger sold would be 2.4 ounces. This is the average of the weight of a regular hamburger (1.6 ounces) and a Big Mac (3.2 ounces). The authors took this approach because

> McDonald's does not publish sales and profits of individual items. Thus, it is not possible to estimate how many of McDonald's first 100 billion beef burgers sold were 1.6 ounce hamburgers, 3.2 ounce Big Macs (introduced in 1968), 4.0 ounce Quarter Pounders (introduced in 1973), or other sandwiches.

This activity can be completed as an individual or as a team. Your instructor will specify which approach (individual or team) you should use.

1. The authors of the article believe that the use of 2.4 ounces as the average size of a burger sold at McDonald's is "conservative," which would result in the estimate of 992.25 million pounds of saturated fat being lower than the actual amount that would be consumed. Explain why the authors' belief might be justified.
2. Do you think it would be possible to collect data that could lead to an estimate of the average burger size that would be better than 2.4 ounces? If so, explain how you would recommend collecting such data. If not, explain why you think it is not possible.

Activity 2.4 Video Games and Pain Management

The purpose of this activity is to provide an opportunity to pose a research question and develop an experiment that would result in data that could be used to answer the question posed.

Background: Video games have been used for pain management by doctors and therapists who believe that the attention required to play a video game can distract the player and thereby decrease the sensation of pain. The paper **"Video Games and Health" (*British Medical Journal* [2005]: 122–123)** states

> However, there has been no long term follow-up and no robust randomized controlled trials of such interventions. Whether patients eventually tire of such games is also unclear. Furthermore, it is not known whether any distracting effect depends simply on concentrating on an interactive task or whether the content of games is also an important factor as there have been no controlled trials comparing video games with other distracters. Further research should examine factors within games such as novelty, users' preferences, and relative levels of challenge and should compare video games with other potentially distracting activities.

1. Working with a partner, select one of the areas of potential research suggested in the passage from the paper and formulate a specific question that could be addressed by performing an experiment.
2. Propose an experiment that would provide data to address the question from Step 1. Be specific about how subjects might be selected, what the experimental conditions (treatments) would be, and what response would be measured.

Copyright 2025 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s). Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

3. At the end of Section 2.3 there are 10 questions that can be used to evaluate an experimental design. Answer these 10 questions for the design proposed in Step 2.
4. After evaluating your proposed design, are there any changes you would like to make to your design? Explain.

Activity 2.5 Be Careful with Random Assignment!

The purpose of this activity is to explore and evaluate different methods for random assignment.

When individuals climb to high altitudes, a condition known as acute mountain sickness (AMS) may occur. AMS is brought about by a combination of reduced air pressure and lower oxygen concentration that occurs at high altitudes. Two standard treatments for AMS are a medication, acetazolamide (which stimulates breathing and reduces mild symptoms) and the use of portable hyperbaric chambers.

With increasing numbers of younger inexperienced mountaineers, it is important to re-evaluate these treatments for the 12- to 14-year age group. An experimental plan under consideration is to study the first 18 youngsters diagnosed with AMS at a high altitude park ranger station whose parents consent to participation in the experiment. Equal numbers of each treatment are desired and the researchers are considering the following strategy for random assignment of treatments: Assign the treatments using a coin flip until one treatment has been assigned nine times; then assign the other treatment to the remaining subjects.

The table below presents data on the first 18 young climbers whose parents consented to participation in the experiment.

Order	Climbing Experience Level	Age
1	beginner	12.90
2	experienced	13.34
3	beginner	12.39
4	beginner	13.95
5	beginner	13.63
6	beginner	13.62
7	experienced	12.55
8	experienced	13.54
9	beginner	12.34
10	experienced	13.74
11	experienced	13.78
12	beginner	14.05
13	experienced	14.22
14	experienced	13.91
15	beginner	14.39
16	experienced	13.54
17	experienced	13.85
18	beginner	14.11

1. Describe how you would implement a strategy equivalent to the one proposed by the researchers. Your plan should assign the treatments M (medicine) and H (hyperbaric chamber) to these climbers as they appear at the ranger station.
2. Implement your strategy in Step (1), assigning treatments to climbers 1–18.
3. Looking at which climbers were assigned to each of the two groups, do you feel that this method worked well? Why or why not?
4. Calculate the proportion of experienced climbers in the medicine group. How does this proportion compare to the proportion of experienced climbers in the entire group of 18 subjects?
5. Construct two dotplots—one of the ages of those assigned to the medicine treatment and one of the ages of those assigned to the hyperbaric chamber treatment. Are the age distributions for the two groups similar?
6. Calculate the average age of those assigned to the medicine group. How does it compare to the average age for the other treatment group?
7. Record the proportion of experienced climbers in the medicine group, the average age of those assigned to the medicine group, and the average age of those assigned to the hyperbaric chamber group obtained by each student in your class.
8. Using the values from Step (7), construct a dotplot of each of the following: the proportion of experienced climbers in the medicine group, the average age of those assigned to the medicine group, and the average age of those assigned to the hyperbaric chamber group.
9. Using the results of the previous steps, evaluate the success of this random assignment strategy. Write a short paragraph explaining to the researchers whether or not they should use the proposed strategy for random assignment and why.

Copyright 2025 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s). Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

Summary Key Concepts and Formulas

Term or Formula	Comment
Observational study	A study that observes characteristics of an existing population.
Simple random sample	A sample selected in a way that gives every different sample of size *n* an equal chance of being selected.
Stratified sampling	Dividing a population into subgroups (strata) and then taking a separate random sample from each stratum.
Cluster sampling	Dividing a population into subgroups (clusters) and forming a sample by randomly selecting clusters and including all individuals or objects in the selected clusters in the sample.
1 in *k* systematic sampling	Forming a sample from an ordered arrangement of a population by choosing a starting point at random from the first *k* individuals on the list and then selecting every *k*th individual thereafter.
Confounding variable	A variable that is related both to group membership and to the response variable.
Measurement or response bias	The tendency for samples to differ from the population because the method of observation tends to produce values that differ from the true value.
Selection bias	The tendency for samples to differ from the population because of systematic exclusion of some part of the population.
Nonresponse bias	The tendency for samples to differ from the population because measurements are not obtained from all individuals selected for inclusion in the sample.
Experiment	A procedure for investigating the effect of *experimental conditions* (treatments) on a *response variable*.
Treatments	The experimental conditions imposed by the experimenter.
Extraneous variable	A variable that is not an explanatory variable in the study but is thought to affect the response variable.
Direct control	Holding extraneous variables constant so that their effects are not confounded with those of the experimental conditions.
Blocking	Using extraneous variables to create groups that are similar with respect to those variables and then assigning treatments at random within each block, thereby filtering out the effect of the blocking variables.
Random assignment	Assigning experimental units to treatments or treatments to trials at random.
Replication	A strategy for ensuring that there is an adequate number of observations for each experimental treatment.
Placebo treatment	A treatment that resembles the other treatments in an experiment in all apparent ways but that has no active ingredients.
Control group	A group that receives no treatment.
Single-blind experiment	An experiment in which the subjects do not know which treatment they received but the individuals measuring the response do know which treatment was received, or an experiment in which the subjects do know which treatment they received but the individuals measuring the response do not know which treatment was received.
Double-blind experiment	An experiment in which neither the subjects nor the individuals who measure the response know which treatment was received.

Chapter Review Exercises 2.69–2.78

2.69 A mortgage lender routinely places advertisements in a local newspaper. The advertisements are of three different types: one focusing on low interest rates, one featuring low fees for first-time buyers, and one appealing to people who may want to refinance their homes. The lender would like to determine which advertisement format is most successful in attracting customers to call for more information.

a. Describe an experiment that would provide the information needed to make this determination. Be sure to consider extraneous variables, such as the day of the week that the advertisement appears in the paper, the section of the paper in which the advertisement appears, or daily fluctuations in the interest rate.

b. What role does random assignment play in your design?

2.70 "Multitasking Increases Study Time, Lowers Grades" is a headline that appeared in UConn Today (July 23, 2015). The article argued that multitasking behaviors like sending texts or checking social media affect college students' academic performance more than they realize and that extra time spent studying is not enough to offset the effect of multitasking. For each of the following study descriptions, answer these five questions.

Question 1: Is the described study an observational study or an experiment?

Question 2: Did the study use random selection from some population?

Question 3: Did the study use random assignment to experimental groups?

Question 4: Would it be reasonable to generalize conclusions from this study to some larger population? If so, what population?

Question 5: Based on this study design, would it be reasonable to conclude that "multitasking lowers grades"?

Study 1: A survey was sent to a random sample of students at a large college. The survey asked students to describe how often they multitasked in class and while doing homework on a scale from "Never" to "Very frequently." The students were also asked to report their current grade point average. The study found that students who self-reported multitasking more frequently had significantly lower grade point averages than those who reported multitasking less frequently.

Study 2: A large university offers tutoring services for students who want to improve their academic performance. Students were selected at random from a list of students who had used these tutoring services, and those selected were invited to participate in a study. The students who agreed to participate consented to be monitored by researchers on two occasions: once while in class and once while studying in the campus library. The researchers recorded the percentage of class/study time participants spent on multitasking behaviors. Students were then asked to report their grades on the next major test or assignment. The study found that students who spent a larger percentage of time on multitasking behavior earned significantly lower grades, on average, than those who spent less time multitasking.

Study 3: A large university offers tutoring services for students who want to improve their academic performance. Students were selected at random from a list of students who had used these tutoring services, and those selected were invited to participate in a study. Students were assigned to one of two groups. Group 1 was asked to leave their possessions including their phones outside the room. Group 2 was allowed to keep their possessions with them. All students were asked to watch a video and read a passage of text introducing new information. Later both groups took the same test to measure how well they could recall what they had learned. The average score for those who had left their possessions outside was significantly higher than the average score of those who had been allowed to keep their phones and other devices that facilitate multitasking with them.

2.71 An article published in the ***Atlanta Journal-Constitution* (July 2, 2021)** included good news and bad news: "The good news is that a Southern diet is delicious and makes us happy. The bad news is that it might raise the risk of our dying early." This statement was based on an observational study of more than 21,000 people ages 45 and older that included 10 years of follow-up. The study found that "participants who ate a Southern-style diet most regularly had a 46% higher risk of sudden cardiac death than people who had the least adherence to this dietary pattern." Based on this study, is it reasonable to conclude that a Southern-style diet causes higher rates of cardiac death? Explain why or why not.

2.72 The article **"I'd Like to Buy a Vowel, Drivers Say" (*USA TODAY*, August 7, 2001)** speculates that young people prefer automobile names that consist of just numbers and/or letters that do not form a word (such as Hyundai's XG300, Mazda's 626, and BMW's 325i). The article goes on to state that Hyundai had planned to identify the car that was eventually marketed as the XG300 with the name Concerto, until they determined that consumers hated it and that they thought XG300 sounded more "technical" and deserving of a higher price. Do the students at your school feel the same way? Describe how you would go about selecting a sample to answer this question.

2.73 An article published in the ***International Journal of Legal Medicine* ("On the Impact of Cannabis Consumption on Traffic Safety: A Driving Simulator Study with Habitual Cannabis Consumers," 2019)** describes an experiment to test how consumption of THC (the main psychoactive compound in marijuana) might affect driving ability. Fifteen participants who were all habitual cannabis users participated in the study. It is clearly unethical to ask study participants to drive after consuming THC, so a driving simulator was used. Each participant took four driving tests using the driving simulator: "sober testing" before consuming any THC, a test

Copyright 2025 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s). Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

immediately after consuming THC, a test 3 hours after consumption, and a test 6 hours after consumption. The experiment did not include a control group or random assignment of THC levels to trials. Explain how this could limit the conclusions drawn from this study.

2.74 An article titled **"The More Pets You Meet as a Baby, the Lower Your Risk of Allergies" (*New Scientist*, December 19, 2018)** reports the results from a study of 1029 children:

> The incidence of allergies was 49 per cent in children who had spent their first 12 months of life in a home with no pet. This fell to 43 per cent in children who as babies had lived with one pet, and 24 per cent for children who had lived with three pets. Two of the children had lived with five pets—neither of them had allergies.

Explain why it is not reasonable to conclude that being raised with more pets is the cause of the observed lower allergy rate.

2.75 A news release from Intel, **"Intel's Security International Internet of Things Smart Home Survey Shows Many Respondents Sharing Personal Data for Money" (March 30, 2016, newsroom.intel.com)**, described a survey conducted in 2015. The news release states "A total of 9,000 consumers were interviewed globally, including 2,500 from the United States, 1,000 from the United Kingdom, 1,000 from France, 1,000 from Germany, 1,000 from Brazil, 1,000 from India, 500 from Canada, 500 from Mexico and 500 from Australia." Among the findings from the survey were that 54% of the respondents worldwide would be willing to share personal data collected from devices in their homes with companies in exchange for money.

Do you think that the study described was an observational study or an experiment?

2.76 In an experiment to assess the effect of wearing compression socks during a marathon, 20 runners in the 2013 Hartford Marathon were randomly assigned to two groups (**"Compression and Clots in Athletes Who Travel," *Lower Extremities Review*, January, 2016)**. Runners in one group wore a pair of compression socks during the marathon, while runners in the second group wore regular athletic socks. At the end of the marathon, blood samples were taken to measure variables related to preventing blood clots and speeding up recovery from exercise.

a. Describe why it was important for the researchers to assign participants to one of the two groups rather than letting the participants choose which group they wanted to be in.

b. The authors of the paper state that there is some evidence that suggests that wearing compression socks may result in a psychological advantage that might translate into performance gains. Suppose that instead of a response variable that was determined by a blood test, the response variable had been the time it took the runner to complete the marathon. Do you think it would be a good idea to have the runners be blind to the type of socks that they were given? Explain why or why not.

2.77 The paper **"Effect of a Nutritional Supplement on Hair Loss in Women" (*Journal of Cosmetic Dermatology* [2015]: 76–82)** describes an experiment to see if a dietary supplement consisting of omega–3, omega–6, and antioxidants could reduce hair loss in subjects with stage 1 hair loss. One hundred twenty subjects volunteered to participate in the study and were randomly assigned to either the supplement group or a control group. The subjects in the supplement group took the supplement for 6 months. Photos of the top of the head were taken of all the subjects at the beginning of the study and 6 months later at the end of the study. The two photos of each subjects were evaluated by an independent expert who visually determined the change in hair density. The expert who determined the change in hair density did not know which of the subjects had taken the supplement.

a. Evaluate this experimental design. Do you think this is a good design or a poor design, and why?

b. If you were designing such a study, what, if anything, would you propose to do differently?

2.78 A manufacturer of clay roofing tiles would like to investigate the effect of clay type on the proportion of tiles that crack in the kiln during firing. Two different types of clay are to be considered. One hundred tiles can be placed in the kiln at any one time. Firing temperature varies slightly at different locations in the kiln, and firing temperature may also affect cracking.

a. Discuss the design of an experiment to collect information that could be used to decide between the two clay types.

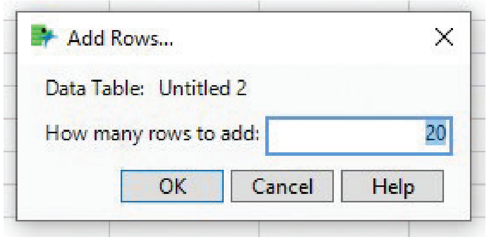
b. How does your proposed design deal with the extraneous variable temperature?

Technology Notes

Generating Random Integers

JMP

1. Once JMP is initialized, select **File > New > DataTable** to create a new data table. A table will appear with "Column 1" in the top row. If you wish, double-click on "Column 1" and change this to "RandSample".
2. Click on Rows > AddRows. The following panel will appear:

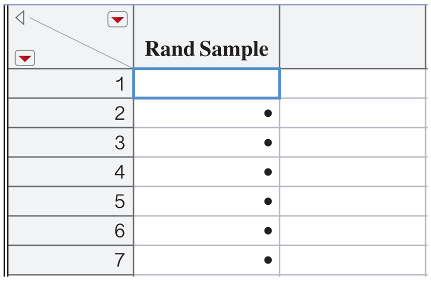

Add Rows...
OK
Cancel
Help

Data Table: Untitled 2

How many rows to add: 20

Source: JMP

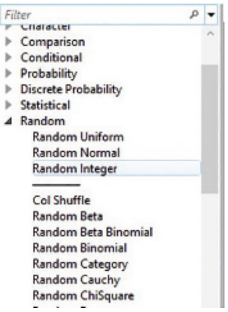
3. Decide how many random integers you wish, and type that number in the box. Then press **OK**. Note that JMP will not protect you against repeats, so you might wish to generate more numbers than you actually need. JMP will then present the column with the desired number of rows.



	Rand Sample	
1		
2	•	
3	•	
4	•	
5	•	
6	•	
7	•	

Source: JMP

4. Right-click on **RandSample** and select **Formula**.
5. Click the triangle next to **Random** to expand the menu.

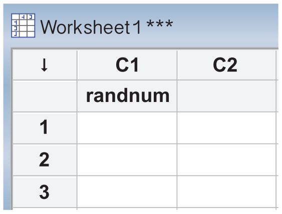

Filter
Character
Comparison
Conditional
Probability
Discrete Probability
Statistical
Random
Random Uniform
Random Normal
Random Integer
Col Shuffle
Random Beta
Random Beta Binomial
Random Binomial
Random Category
Random Cauchy
Random ChiSquare

Source: JMP

6. Click on **Random Integer**. The "Formula" box will update to "RandomInteger (k)".
7. Double click on k then replace k with the maximum value desired and press **ENTER**. (If you want just random integers from 1 to 50, you substitute 50 for k.) The random integers will appear in the RandSample column after clicking **OK**.

Minitab

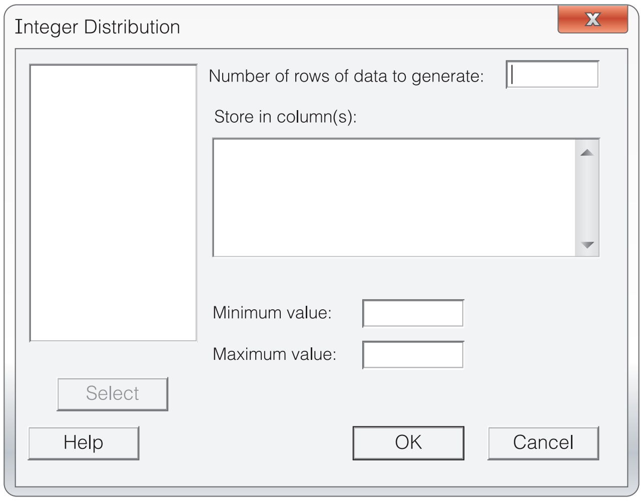
1. In the main Minitab Data Editor, at the top of the C1 variable column, enter "randnum" as the variable name.


Worksheet 1 ***

↓	C1	C2
	randnum	
1		
2		
3		

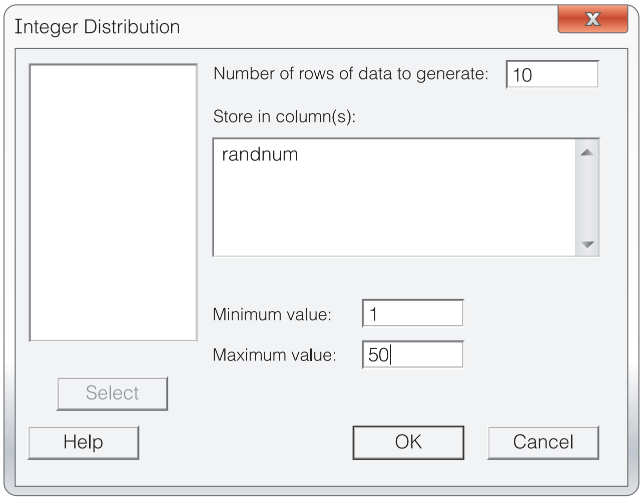
Source: Minitab

2. In succession, click on **Calc**, **RandomData**, and finally **Integer**. The Integer Distribution panel will appear.


Integer Distribution
Number of rows of data to generate:
Store in column(s):
Minimum value:
Maximum value:
Select
Help
OK
Cancel

Source: Minitab

3. Click on the **Store in column(s)** panel, and then double-click on **C1 randnum**.


Integer Distribution
Select
Help
OK
Cancel

Number of rows of data to generate: 10

Store in column(s): randnum

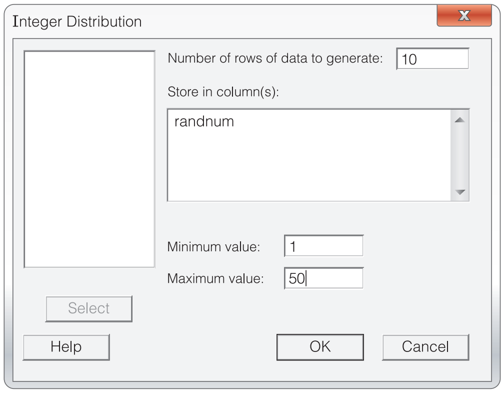
Minimum value: 1

Maximum value: 50

Source: Minitab

4. Enter the number of random integers desired and the minimum and maximum values.

Copyright 2025 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s). Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.


Integer Distribution
Select
Help
OK
Cancel

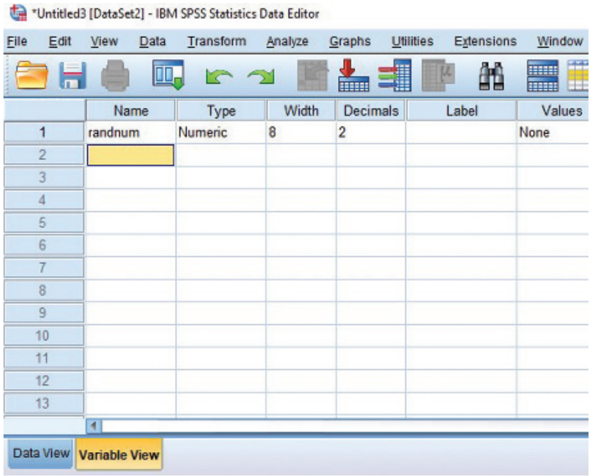
Number of rows of data to generate: 10
Store in column(s): randnum
Minimum value: 1
Maximum value: 50

Source: Minitab

5. Press OK. The random integers will appear in the randnum column.

SPSS

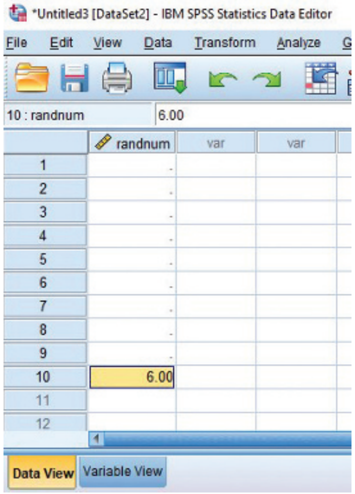
1. In the SPSS Data Editor, click on the Variable View tab at the bottom of the screen. Then enter randnum in the name column and hit return.


*Untitled3 [DataSet2] - IBM SPSS Statistics Data Editor
File Edit View Data Transform Analyze Graphs Utilities Extensions Window
Data View
Variable View

	Name	Type	Width	Decimals	Label	Values
1	randnum	Numeric	8	2		None

Source: SPSS

2. Click on the Data View tab at the bottom of the screen. Then enter any number in the cell corresponding to how many random numbers you want to generate. For example, to generate 10 random numbers, type any number in cell 10 of the randnum column.

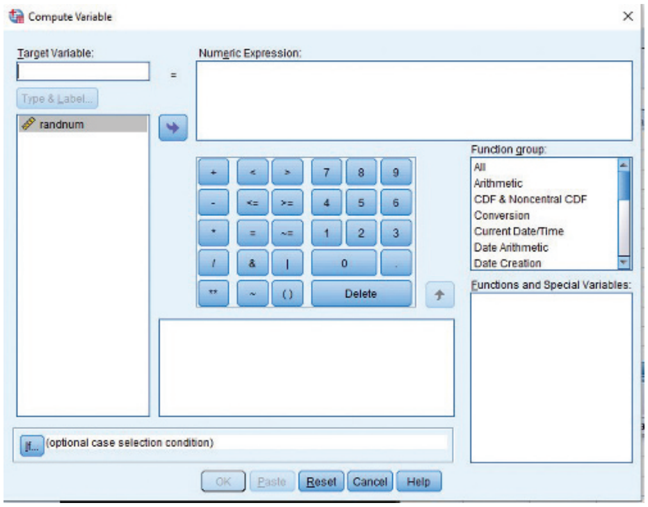

*Untitled3 [DataSet2] - IBM SPSS Statistics Data Editor
File Edit View Data Transform Analyze G
var
var
Data View
Variable View

10 : randnum 6.00

	randnum
1	.
2	.
3	.
4	.
5	.
6	.
7	.
8	.
9	.
10	6.00
11	
12	

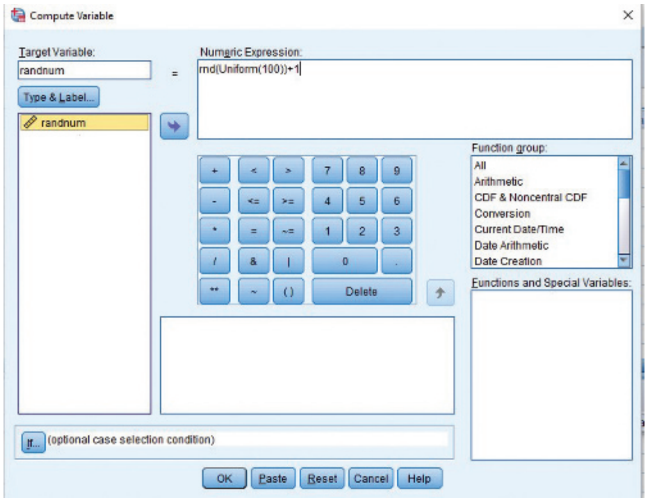
Source: SPSS

3. Choose Compute Variable from the Transform menu.


Compute Variable
Target Variable:
Numeric Expression:
Type & Label...
randnum
Function group:
All
Arithmetic
CDF & Noncentral CDF
Conversion
Current Date/Time
Date Arithmetic
Date Creation
Functions and Special Variables:
Delete
If... (optional case selection condition)
OK
Paste
Reset
Cancel
Help

Source: SPSS

4. Type randnum in the Target Variable box. Type RND(Uniform(N))+1 in the Numeric Expression box. N is a number that represents the largest integer you want to include. For example, to generate random integers between 1 and 100, you would type 100 for N.


Compute Variable
Type & Label...
randnum
Function group:
All
Arithmetic
CDF & Noncentral CDF
Conversion
Current Date/Time
Date Arithmetic
Date Creation
Functions and Special Variables:
Delete
If... (optional case selection condition)
OK
Paste
Reset
Cancel
Help

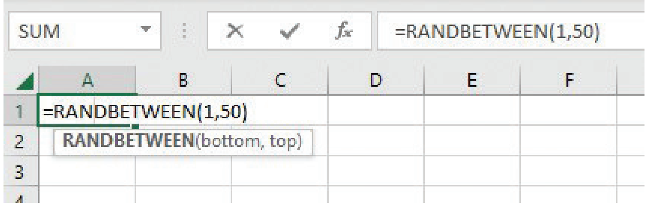
Target Variable: randnum
Numeric Expression: rnd(Uniform(100))+1

Source: SPSS

5. Click on OK and then click on OK in the box that asks if you want to change the existing variable. The requested random numbers will now appear in the randnum column of the Data Editor.

Excel

1. Choose a cell in Excel and enter "=RANDBETWEEN (bottom, top)" where bottom and top are the lower and upper ends of the range of values of random integers that you want. Press **ENTER**.

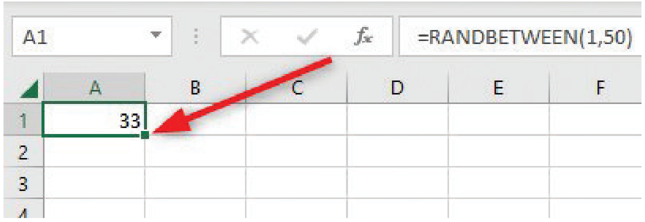

SUM
A
B
C
D
E
F
RANDBETWEEN(bottom, top)

=RANDBETWEEN(1,50)

Source: Microsoft Corporation

Copyright 2025 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s). Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

2. Click on the small square in the lower right of the cell. Drag the square downward the number of rows equal to the number of random integers that you wish to have.



A1

=RANDBETWEEN(1,50)

33

Source: Microsoft Corporation

3. Check for any repetitions. If found, use the lower-right-corner technique to add more.

TI-83/84

Note: The TI calculator has a built-in sequence of pseudo-random numbers that has been determined by "seeding" the starting random number. Unless this number is changed, everyone in a group with a new calculator will get the same sequence of random numbers. Because of this, Step 0 is the "reseeding" step. Normally, this would only be done one time.

0. At the home screen, enter an 8-digit integer. Press **STO > MATH > PROB**.
Choose "**rand**" and press **ENTER**.

To generate n random numbers, with no repeats, between 1 and N:

1a. Select **MATH > PROB > randIntNoRep(**
2a. Supply the relevant information about the desired random integers. For example,
lower: 1
upper: 50
n: 7
Paste

Press **ENTER** after each of these lines.

3a. **randIntNoRep(1, 50, 7)** will appear. Press **ENTER** and n random integers will be displayed.

If your calculator does not have the most recent operating system, you must supply the numbers indicated. The randInt function will not necessarily give nonrepeating integers.

1b. Select **MATH > PROB > randInt(**
2b. Enter **1**, **50**, **7**, ")" and **ENTER**.
3b. **randInt(1, 50, 7)** should appear on the screen. Press **ENTER** and n = 7 numbers will be displayed.

To store the random integers in a TI List, do not press ENTER after **randIntNoRep(1, 50, 7)** or **randInt(1, 50, 7)** in Step 3. Instead, press **STO> 2nd > L1 > ENTER**. The random integers will be stored in List1.

TI-Nspire

Note: The TI calculator has a built-in sequence of pseudo-random numbers that has been determined by "seeding" the starting random number. Unless this number is changed, everyone in a group with a new calculator will get the same sequence of random numbers. Because of this, Step 0 is the "reseeding" step. Normally, this would be only be done one time.

0. Open a new document. Press **Menu** then select **1: Add Calculator**. Press **Menu** and select **5: Probability** then **4. Random** and **6: Seed**. Type an 8-digit integer and press **enter**.

To generate n random numbers, with no repeats, between 1 and N:
1. Press **Menu** and select **5: Probability** then **4: Random** and **2: Integer**.
3. **randInt()** will appear. To generate 7 random integers between 1 and 50 type 1, 50, 7, ")" and press **enter**. Seven random integers between 1 and 50 will be displayed. If there are repeats, press **enter** again to generate another set of random integers.

Chapter 3

Graphical Methods for Describing Data

Africa Studio/Shutterstock.com

Most college students are concerned about the cost of a college education. **The National Center for Education Statistics (nces.ed.gov)** reported the average in state tuition and fees for 4-year public institutions in each of the 50 U.S. states for the 2021–2022 academic year. Average tuition and fees (in dollars) are given below for each state.

Several questions could be posed about these data. What is a typical value of average in-state tuition and fees for the 50 states? Are observations concentrated near the typical value, or do average tuition and fees differ quite a bit from state to state? Are there any states whose average tuition and fees are unusual compared to the rest? What proportion of the states have average tuition and fees less than $7500? More than $10,000?

Questions like these are most easily answered if the data can be organized in a sensible manner. In this chapter, some techniques for organizing and describing data using tables and graphs are introduced.

7,861	7,324	5,037	4,554	3,148	7,106	8,741	7,643	3,795	4,737
5,020	6,259	6,155	10,304	6,709	5,282	6,496	6,772	8,732	8,403
7,142	8,964	7,238	7,551	5,004	4,731	3,542	6,547	5,904	10,084
10,103	3,653	5,395	7,546	7,152	6,285	7,702	12,190	10,338	8,601
7,971	6,825	6,476	6,085	8,537	13,448	5,995	6,308	6,424	4,429

Learning Objectives

After completing this chapter, you will be able to:

- Explain how selecting an appropriate graphical display depends on the data type (categorical or numerical) and whether or not the purpose of the display is to compare groups.
- Describe a numerical data distribution in terms of center, shape, and spread.
- Use a scatterplot to investigate the relationship between two numerical variables.
- Use a time series plot to investigate a trend over time.
- Construct and interpret graphical displays of categorical data: pie charts and segmented bar charts.

- Construct and interpret graphical displays of numerical data: stem-and-leaf displays, histograms, and relative frequency histograms.
- Construct and interpret graphical displays designed to compare groups: comparative bar charts and comparative stem-and-leaf displays.
- Construct and interpret a scatterplot of bivariate numerical data.
- Construct and interpret a time series plot.
- Describe relationships between three or more variables based on multivariable graphical displays.

Section 3.1 Displaying Categorical Data: Comparative Bar Charts and Pie Charts

Comparative Bar Charts

In Chapter 1 you saw that categorical data could be summarized in a frequency distribution and displayed graphically using a bar chart. Bar charts can also be used to give a visual comparison of two or more groups. This is accomplished by constructing two or more bar charts that use the same set of horizontal and vertical axes, as illustrated in Example 3.1.

Example 3.1 Is Education Worth the Cost?

Understand the context ➤

Do people with college degrees think that their education was worth the cost? This question was posed to 2548 adults with an associate degree and to 30,151 adults with a bachelor's degree. The data, from the Gallup report **"Two-Year Grads Satisfied with Cost of Degree" (gallup.com, April 11, 2016)**, are summarized in the accompanying table.

Consider the data ➤

	Frequency		Relative Frequency	
Response	**Associate Degree Holders**	**Bachelor's Degree Holders**	**Associate Degree Holders**	**Bachelor's Degree Holders**
Strongly disagree	178	1,508	0.07	0.05
Disagree	153	2,111	0.06	0.07
Neither agree nor disagree	408	4,221	0.16	0.14
Agree	637	8,743	0.25	0.29
Strongly agree	1,172	13,568	0.46	0.45
Total	**2,548**	**30,151**	**1.00**	**1.00**

What Do You Think?

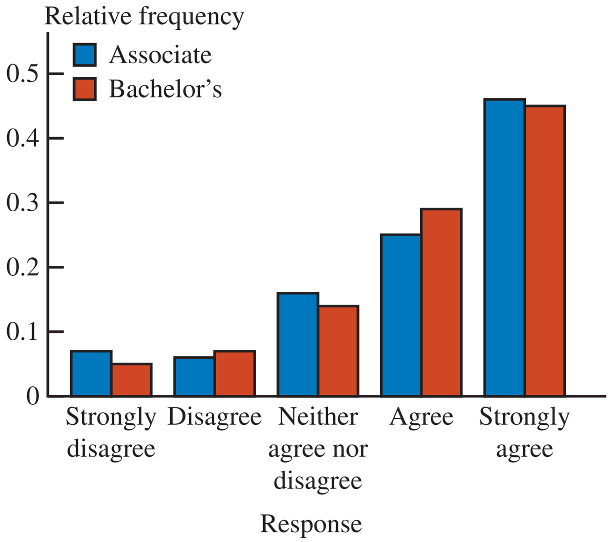
If you wanted to compare the responses of associate degree holders and the responses of bachelor's degree holders, would you use the response frequencies or the response relative frequencies?

Because the two sample sizes are very different, it is important to use relative frequencies rather than frequencies to construct the scale for the vertical axis in the comparative bar chart. The same set of steps that were used to construct a bar chart are used to construct a comparative bar chart, but in a comparative bar chart each category will have bars for all of the groups.

Interpret the results ➤

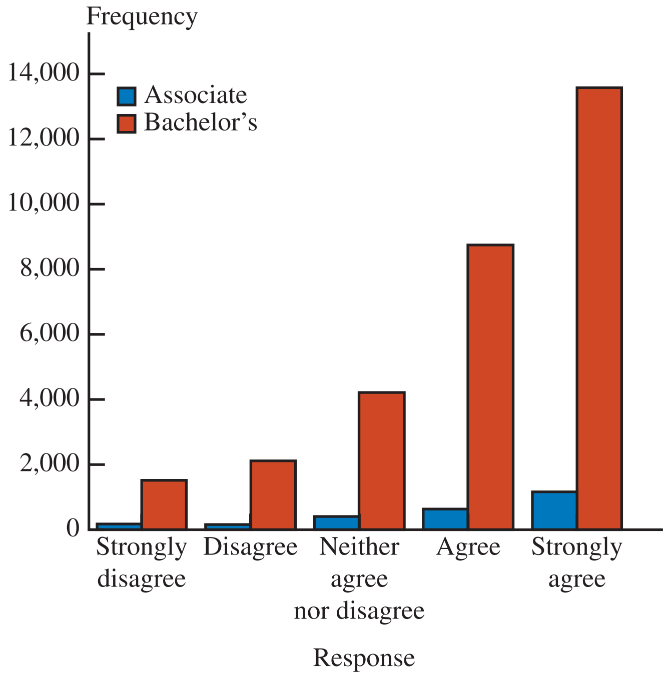
The comparative bar chart for these data is shown in Figure 3.1. Looking at the comparative bar chart, it is easy to compare the response distributions of the two groups. Notice that the response distributions are similar, indicating that both associate degree holders and bachelor's degree holders generally agreed that their education was worth the cost, with more than 70% in the Agree or Strongly agree categories.

Figure 3.1
Comparative bar chart.



To see why relative frequencies are used to compare groups of different sizes, consider the *incorrect* comparative bar chart constructed using the frequencies rather than the relative frequencies (Figure 3.2). The incorrect comparative bar chart conveys a very different and misleading impression.

Figure 3.2
An **incorrect** comparative bar chart for the data of Example 3.1.



When constructing a comparative bar chart, use relative frequency rather than frequency to construct the scale on the vertical axis so that meaningful comparisons can be made even if the sample sizes are not the same.

Pie Charts

A categorical data set can also be summarized using a pie chart. In a pie chart, a circle is used to represent the whole data set, with "slices" of the pie representing the possible categories. The size of the slice for a particular category is proportional to the corresponding frequency or relative frequency. Pie charts are most effective for summarizing data sets when there are not too many different categories.

Example 3.2 Life Insurance for Cartoon Characters?

Understanding the context ➤

The article **"Fred Flintstone, Check Your Policy" (*The Washington Post,* October 2, 2005)** summarized the results of a survey of 1014 adults conducted by the Life and Health Insurance Foundation for Education. Each person surveyed was asked to select which of five fictional characters, Spider-Man, Batman, Fred Flintstone, Harry Potter, and Marge Simpson, they thought had the greatest need for life insurance. The resulting data are summarized in the pie chart of Figure 3.3.

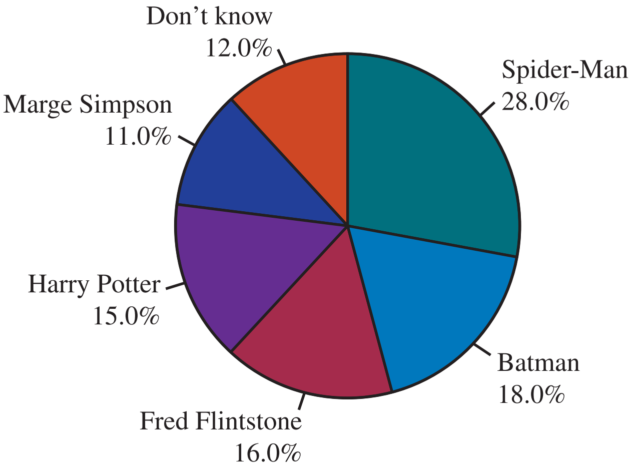


Figure 3.3
Pie chart of data on which fictional character most needs life insurance.

Interpret the results ➤

The survey results were quite different from an insurance expert's assessment. The expert's opinion was that Fred Flintstone, a married father with a young child, was by far the one with the greatest need for life insurance. Spider-Man, unmarried with an elderly aunt, would need life insurance only if his aunt relied on him to supplement her income. Batman, a wealthy bachelor with no dependents, doesn't need life insurance in spite of his dangerous job!

Pie Chart for Categorical Data

When to Use Categorical data with a relatively small number of possible categories. Pie charts are useful for illustrating proportions of the whole data set for various categories.

How to Construct

1. Draw a circle to represent the entire data set.
2. For each category, calculate the "slice" size. Because there are 360 degrees in a circle

   slice size = 360·(category relative frequency)

3. Draw a slice of appropriate size for each category. This can be tricky, so most pie charts are generated using a graphing calculator or a statistical software package.

What to Look For

- Categories that represent particularly large and small proportions of the data set.

Example 3.3 Watch Those Typos

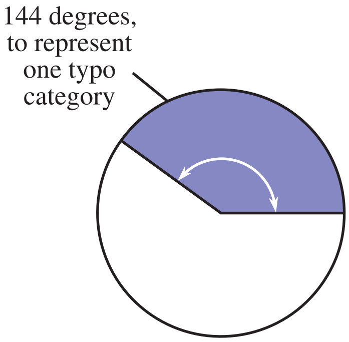
Understanding the context ➤

Typos on a résumé do not make a very good impression when applying for a job. Senior executives were asked how many typos in a résumé would make them not consider a job candidate (**"Job Seekers Need a Keen Eye," *USA TODAY,* August 3, 2009**). The resulting data are summarized in the accompanying relative frequency distribution.

Consider the data ➤

Number of Typos	Frequency	Relative Frequency
1	60	0.40
2	54	0.36
3	21	0.14
4 or more	10	0.07
Don't know	5	0.03

Do the work ➤



To draw a pie chart by hand, you would first compute the slice size for each category. For the one typo category, the slice size would be

$$\text{slice size} = (0.40)(360) = 144 \text{ degrees}$$

You would then draw a circle and use a protractor to mark off a slice corresponding to about 144°, as illustrated in the figure shown in the margin. Continuing to add slices in this way leads to a completed pie chart.

It is much easier to use a statistical software package to create pie charts than to construct them by hand. A pie chart for the typo data, created with the statistical software package Minitab, is shown in Figure 3.4.

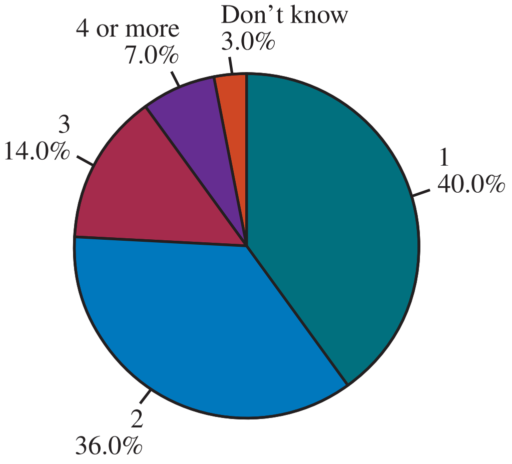


Figure 3.4
Pie chart for the typo data of Example 3.3.

Interpret the results ➤

From the completed pie chart, it is easy to see that even one or two typos would result in many employers not considering a candidate for a job.

Pie charts can be used effectively to summarize a single categorical data set if there are not too many different categories. However, pie charts are not usually the best tool if the goal is to compare groups on the basis of a categorical variable. This is illustrated in Example 3.4.

Example 3.4 Teens and Social Media

Understand the context ➤

Researchers at the Pew Research Center (pewresearch.org) carried out a survey on social media use. They asked 504 teens age 13 to 14 and 812 teens age 15 to 17 if they thought the amount of time that they spend on social media is too much, about right, or too little. The resulting data were used to create the two pie charts in Figure 3.5.

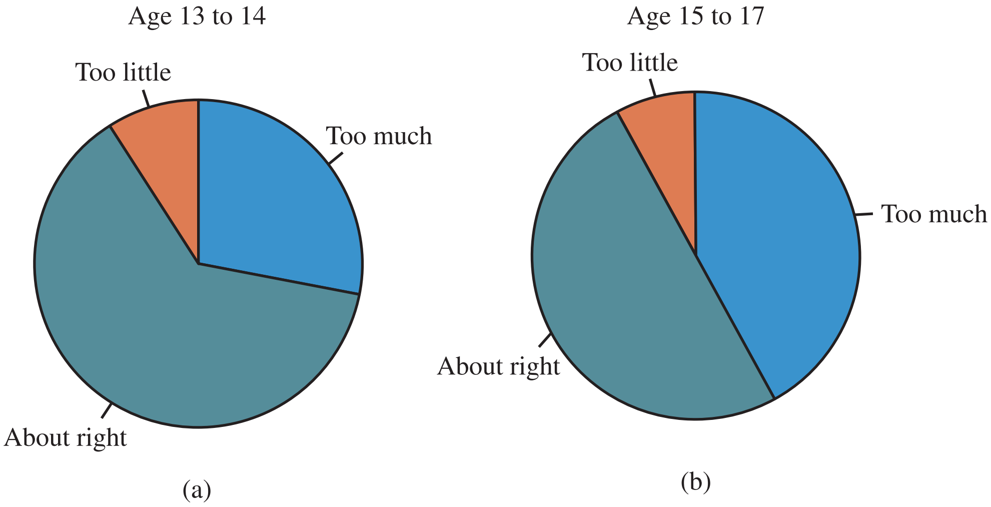


Figure 3.5
Pie charts for Example 3.4:
(a) age 13 to 14:
(b) age 15 to 17.

Although the differences in responses for the two age groups can be seen by comparing the pie charts of Figure 3.5, it can sometimes be difficult to compare category proportions using pie charts. A comparative bar chart (Figure 3.6) makes this type of comparison easier.

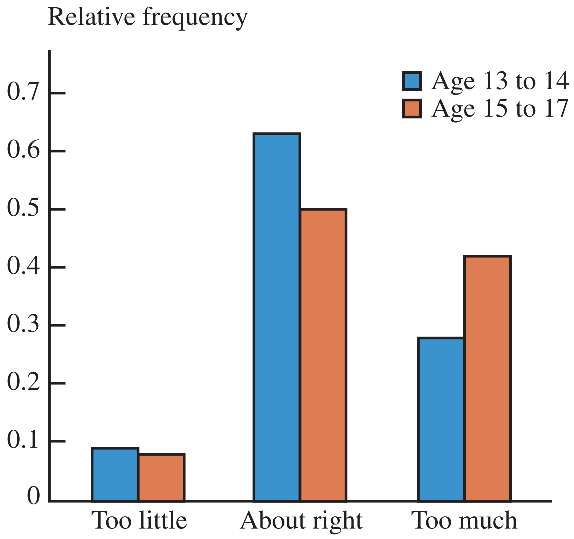


Figure 3.6
Comparative bar chart of responses for the two age groups.

Interpret the results ➤ From the comparative bar chart, it is easy to see that the teens age 15 to 17 were much more likely to say that they spend too much time on social media than the younger teens and much less likely to say that they spend about the right amount of time.

A Different Type of “Pie” Chart: Segmented Bar Charts

A pie chart can be difficult to construct by hand, and the circular shape sometimes makes it difficult to compare areas for different categories, particularly when the relative frequencies for categories are similar. The **segmented bar chart** (also sometimes called a stacked bar chart) avoids these difficulties by using a rectangular bar rather than a circle to represent the entire data set. The bar is divided into segments, with different segments representing different categories. As with pie charts, the area of the segment for a particular category is proportional to the relative frequency for that category. Example 3.5 illustrates the construction of a segmented bar chart.

Copyright 2025 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s). Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

Example 3.5 Cheating in Online Courses

Understand the context ➤

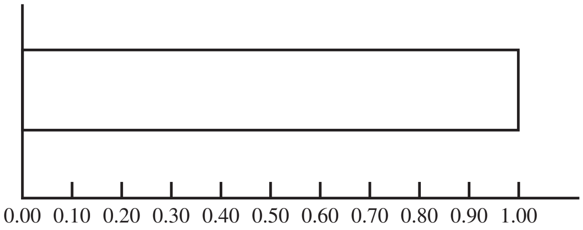
In 2020, most college students were taking courses online due to the COVID-19 pandemic. The report **"Student Success Beyond COVID-19" (Inside Higher Ed and College Pulse, 2021)** summarized data from a survey of 2000 college students conducted in May 2021. The accompanying relative frequency table summarizes the responses to the question, "To your knowledge, how prevalent is cheating in the online courses you are taking?"

Consider the data ➤

Response	**Relative Frequency**
Extremely common	0.21
Somewhat common	0.26
Somewhat uncommon	0.10
Extremely uncommon	0.06
I'm not sure	0.35
I'm not taking any online courses	0.02

Do the work ➤

To construct a segmented bar chart for these data, first draw a bar of any fixed width and length, and then add a scale that ranges from 0 to 1, as shown.



Then divide the bar into six segments, corresponding to the six possible response categories in this example. The first segment, corresponding to the extremely common category, ranges from 0 to 0.21. The second segment, corresponding to somewhat common, ranges from 0.21 to 0.47 (for a length of 0.26, the relative frequency for this category), and so on. The segmented bar chart is shown in Figure 3.7.

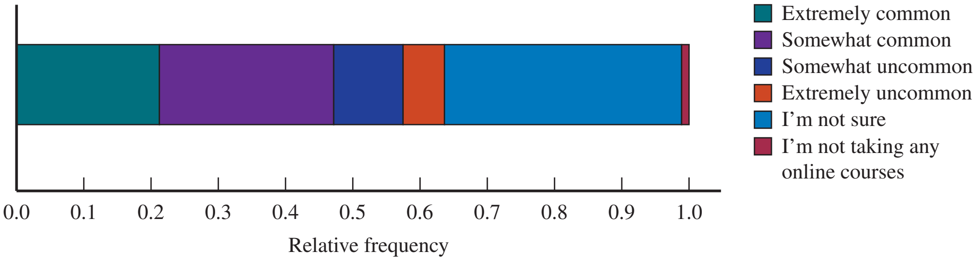


Figure 3.7
Segmented bar chart for the cheating prevalence data of Example 3.5.

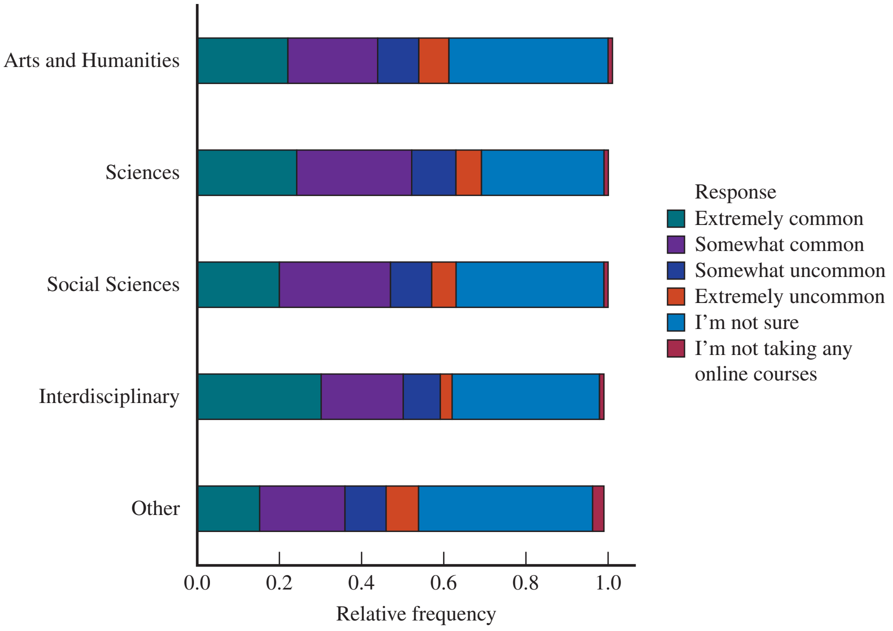
The same report also gave data on the responses to this question by college major and included a display similar to the one in Figure 3.8. Viewing these displays side by side makes it easy to see differences in the response distributions for the different majors.

What Do You Think?

Looking at Figure 3.8, in what ways are the response distributions different for the different majors? In what ways are they similar?

Copyright 2025 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s). Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

Figure 3.8
Segmented bar charts for cheating prevalence data by major.



Interpret the results ➤ There are differences in the response distributions for the different majors, with the biggest differences in the extremely common and not sure responses.

Other Uses of Bar Charts and Pie Charts

As you have seen in previous examples, bar charts and pie charts can be used to summarize categorical data sets. However, they are sometimes also used for other purposes, as illustrated in Examples 3.6 and 3.7.

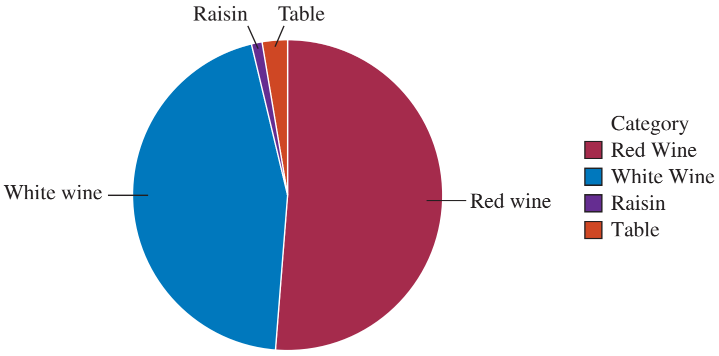
Example 3.6 Grape Production

● The 2020 **Grape Crush Report for California** gave the following information on grape production for each of four different types of grapes **(nass.usda.gov)**:

Type of Grape	Tons Produced
Red Wine Grapes	1,814,000
White Wine Grapes	1,590,000
Raisin Grapes	42,000
Table Grapes	95,000
Total	**3,541,000**

Although this table is not a frequency distribution, it is common to represent information of this type graphically using a pie chart, as shown in Figure 3.9. The pie represents the total grape production, and the slices show the proportion of the total production for each of the four types of grapes.

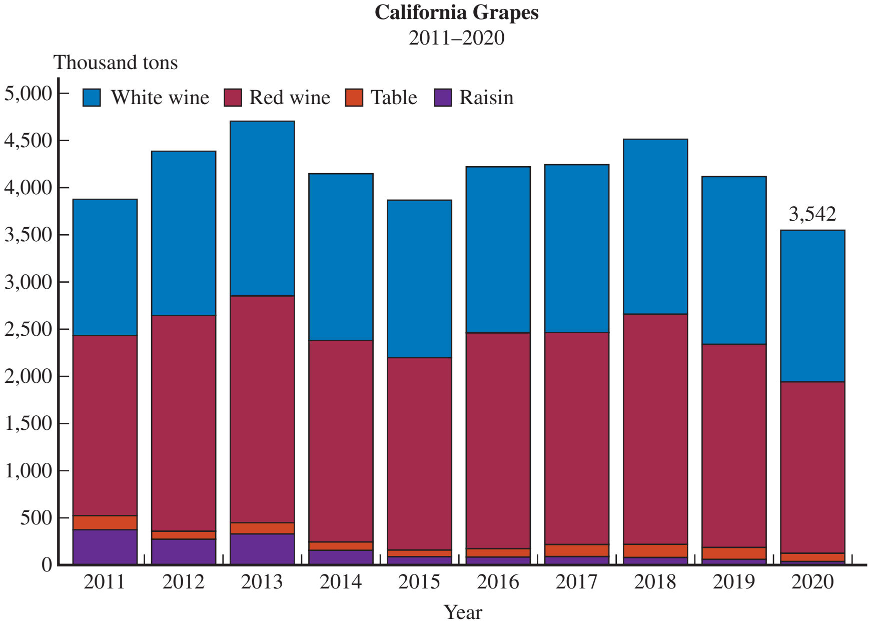
Figure 3.9
Pie chart for grape production data.



● **Data set available online**

The 2020 report also included segmented bar charts like the ones shown in Figure 3.10 that represent how grape production has changed over time. It is easier to see the change over time from these segmented bar charts that can be displayed side by side than it would be from 10 pie charts displaying the same information. From the segmented bar charts, you can see that total grape production is down a bit from its high in 2013. There has been a noticeable decrease in the production of raisin grapes, and the production of all four types of grapes was down in 2020.

Figure 3.10
Segmented bar charts showing change in grape production over time.



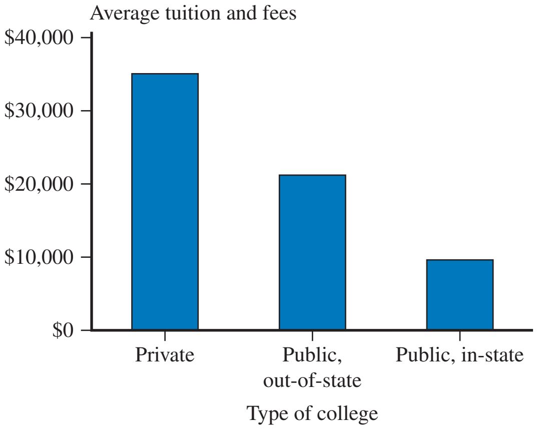
Example 3.7 College Tuition

● *U.S. News & World Report* reported the average tuition and fees for public and private four-year colleges in 2020 (**www.usnews.com**), shown in the accompanying table.

Type of College	Average Tuition and Fees
Private	$35,087
Public, out-of-state	$21,184
Public, in-state	$9,687

Even though this table is not a frequency distribution, this type of information is often represented graphically in the form of a bar chart, as illustrated in Figure 3.11. In this type of bar chart, the height of the bar represents the value of a numerical variable.

Figure 3.11
Bar chart of average tuition and fees in 2020.

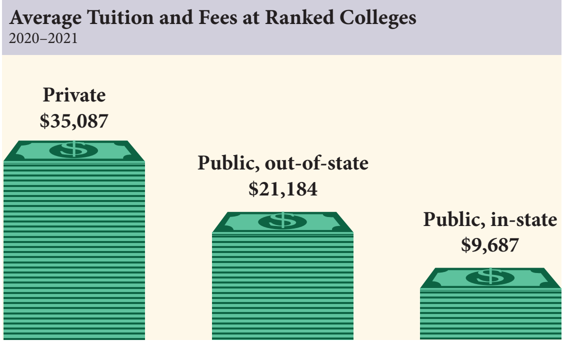


● **Data set available online**

Copyright 2025 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s). Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

A graph similar to one appearing in the *U.S. News & World Report* article is shown in Figure 3.12. As is typical of infographics appearing in the media, they have made the bar chart more interesting by making the bar appear like a stack of dollar bills.

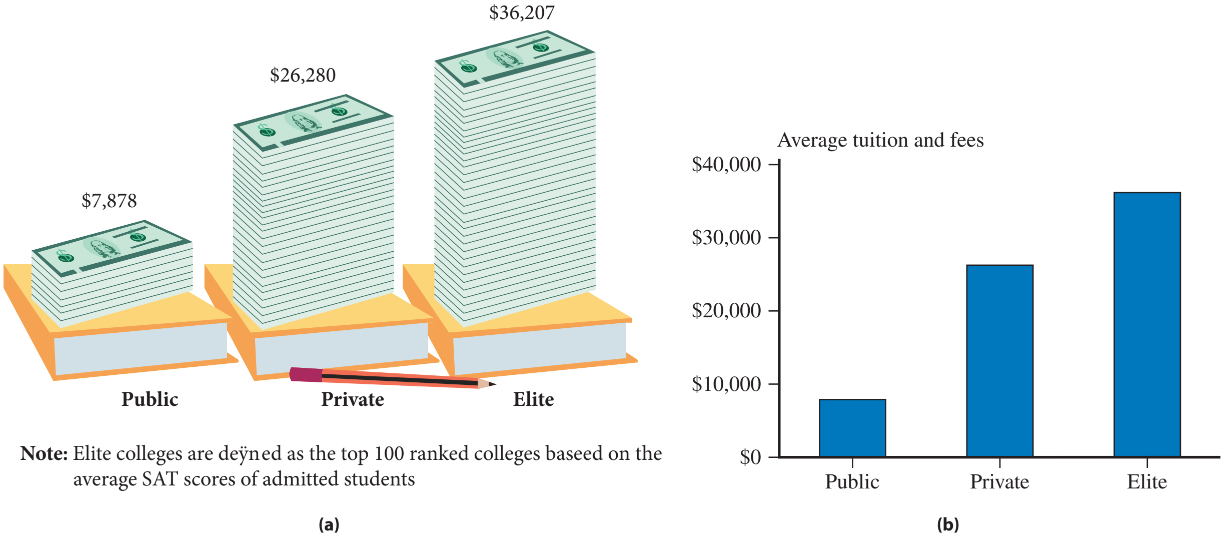
Figure 3.12
Bar chart of average tuition in fees in 2020 from media report.



Both of the bar charts shown in Figures 3.11 and 3.12 allow you to see the relative differences in average tuition and fees for the different types of colleges.

In 2005, *USA Today* published an infographic that showed the predicted average tuition and fees for 2020. Figure 3.13(a) is similar to the infographic that appeared in *USA Today*. Although the bar chart is meant to represent the averages, they have added a third dimension to the bar chart that makes it a bit misleading in terms of comparing the different types of colleges because the areas of the "bars" are not proportional to the values that the bars represent. Figure 3.13(b) gives a two-dimensional bar chart that allows for more accurate comparisons.

Figure 3.13
Bar charts for predicted 2020 average tuition (a) Misleading representation from the media (b) Correctly drawn bar chart.



Note: Elite colleges are defined as the top 100 ranked colleges baseed on the average SAT scores of admitted students

(a)

(b)

What Do You Think?

How did the predictions of average tuition and fees made in 2005 (see Figure 3.13) compare to the actual averages (see Figures 3.11 and 3.12)?

Copyright 2025 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s). Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

Exercises 3.1–3.14

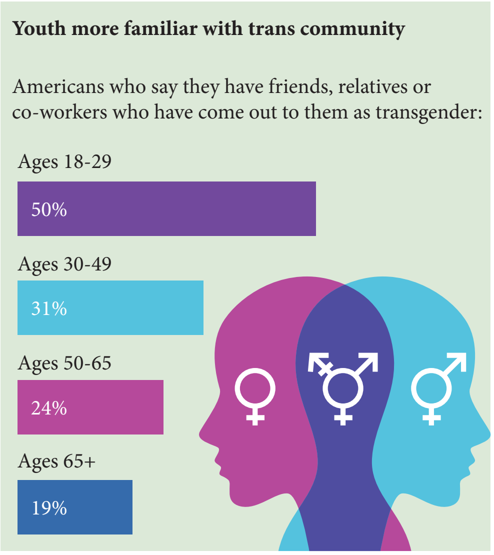
● **Data set available online**

3.1 Each person in a nationally representative sample of people who work for companies that allow employees to work with a combination of both on-site work and work from home (called a hybrid workplace) was asked about the hybrid work policies of their employers **("Coordinating Hybrid Work Schedules—5 Important Findings," gallup.com, November 7, 2022)**. The accompanying table summarizes the responses.

Response	**Relative Frequency**
Employer requires on-site attendance on specific days of the week	0.29
Employer requires a minimum number of days on-site but doesn't specify which days are required	0.28
No universal on-site requirements	0.43

Summarize this information in a pie chart. (Hint: See Examples 3.2 and 3.3.)

3.2 The accompanying graphical display is similar to one that appeared in ***USA TODAY* (June 22, 2021).** It is a bar graph of the percentages of Americans who say they have friends, relatives, or co-workers who have come out to them as transgender for each of four age groups.

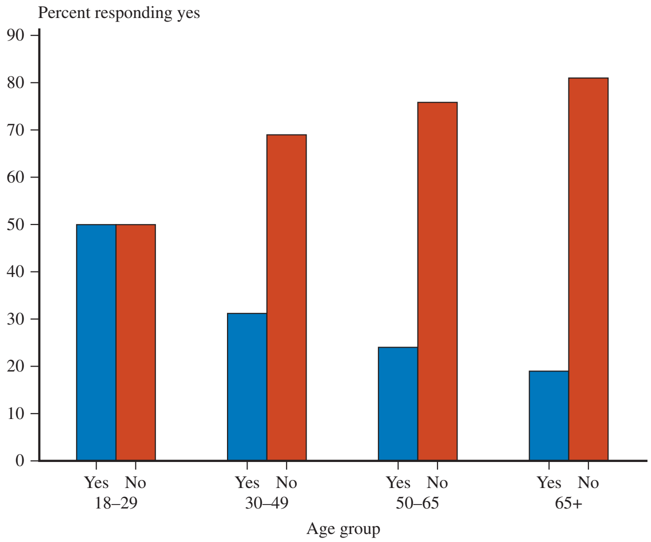


Source: Gallup survey, May 3–18

a. The data used to calculate the percentages in the bar chart were based on responses to the following question: Have you had a friend, relative or co-worker who has come out to you as transgender? Would the responses to this question be categorical or numerical?

b. Explain why a bar chart rather than a dotplot was used to summarize the response data.

c. These data could have also been displayed using the following comparative bar chart. Which of the two displays do you think is more effective for communicating the survey results? Explain.



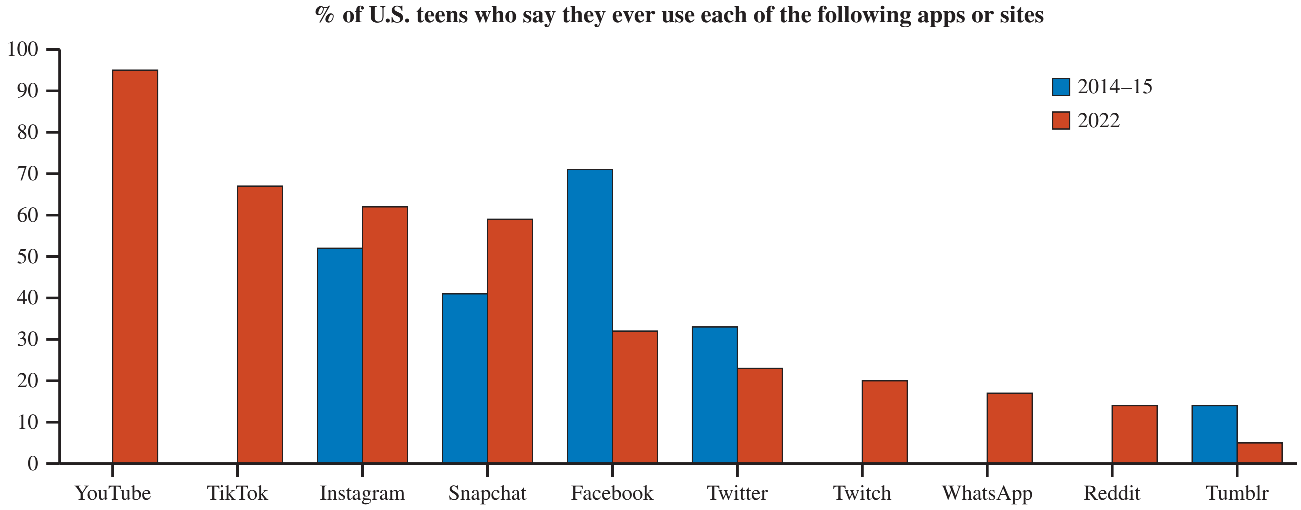
d. The title in the graph that appeared in *USA Today* was "Youth more familiar with trans community." How is this conclusion supported by what you see in the bar chart and the comparative bar chart?

3.3 The comparative bar chart at the top of the next page is similar to ones appearing in the report **"Teens, Social Media and Technology 2022" (Pew Research Center, August 2022)**. The bar charts summarize responses from large representative samples of teens age 13 to 17 in surveys conducted in 2014–2015 and 2022. Notice that some of the apps included in the 2022 survey were not included in the 2014–2015 survey.

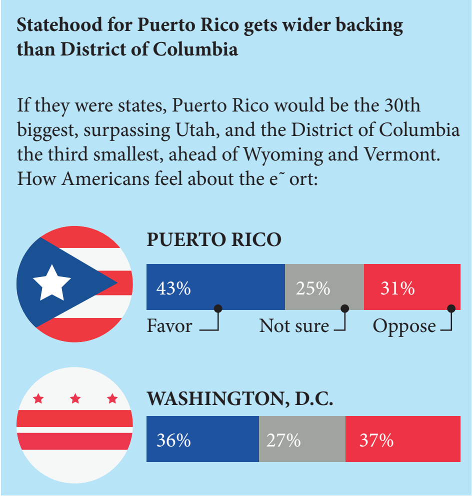
Suppose that you plan to include these displays in an article that you are writing for your school newspaper. Write a few sentences that could accompany the displays. Be sure to comment on any notable differences between 2014–2015 and 2022.

Copyright 2025 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s). Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

Figure for Exercise 3.3



3.4 The accompanying graphical display is similar to one that appeared in ***USA Today* (June 14, 2021)**. It summarizes survey responses to a question about statehood for Puerto Rico and Washington D.C. Which of the graph types introduced in this section is used here? (*USA TODAY* frequently adds artwork and text to their graphs to make them look more interesting.)

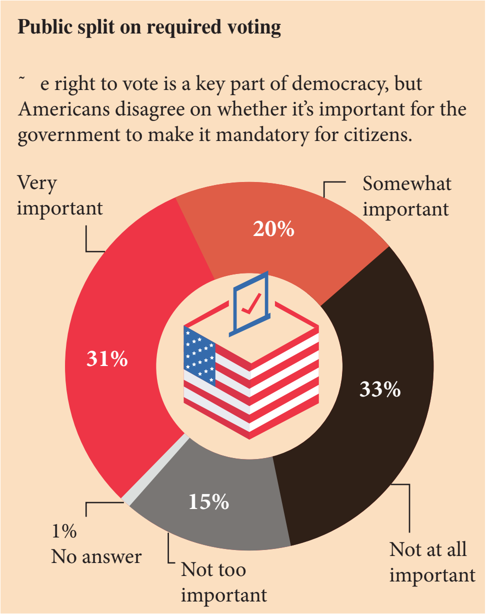


Source: U.S. Census Bureau; The Economist/YouGov poll.

3.5 The accompanying graph is similar to one that appeared in ***USA TODAY* (June 15, 2021)**. This graph is a modified pie chart.

a. Use the information in the graph to construct a traditional pie chart and a bar chart.

b. Which of the three displays do you think is more effective for communicating the survey results? Explain.



Source: Pew Research Center

3.6 ● *Consumer Reports* evaluates products and assigns both an overall score and a grade (there are five grades ranging from worst to best). The scores and grades for 20 fitness trackers are summarized in the table at the top of the next page **(www.consumerreports.org)**.

a. Two variables are summarized in the table: grade and overall score. Is overall score a numerical or categorical variable? Is grade (indicated by the different colors in the table) a numerical or categorical variable?

WORST **BEST**

Fitness Tracker	**Overall Score**
Garmin Forerunner 235	85
Fitbit Inspire 2	84
Garmin Forerunner 35	84
Samsung Galaxy Fit2	80
Fitbit Charge 3	76
Samsung Galaxy Fit	76
Fitbit Charge 4	75
Casio G-Shock Move	74
Fitbit Inspire HR	73
Garmin Vivosmart 4	73
Withings Pulse HR	71
Polar Ignite 2	71
Garmin Vivofit 4	71
Fitbit Inspire	69
Garmin Forerunner 245 Music	64
Xiaomi Mi Band 5	62
Garmin Vivofit jr. 2	56
Polar Grit X	52
Xiaomi Mi Band 4	38
Fitbit Ace 2	35

b. Explain how the accompanying table is equivalent to a segmented bar chart of the grade data.

c. Construct a dotplot of the overall score data. Based on the dotplot, suggest an alternate assignment of grades to the 20 fitness trackers. Explain the reasoning you used to make your assignment.

3.7 The graph at the bottom of the page is similar to one that appeared in the article **"The Challenge of Cutting the Federal Budget, Visualized" (*The Washington Post*, January 26, 2023).**

a. In what way is this graph similar to a pie chart?

b. Do you think a pie chart would have been as effective in summarizing how the federal government spends money? Explain your answer.

3.8 The graph at the top of the next page is similar to one that appeared in the report **"Partisans in the U.S. Increasingly Divided on Whether Offensive Content Online Is Taken Seriously Enough" (Pew Research Center, October 8, 2020).** Survey participants were asked to choose between the following two statements:

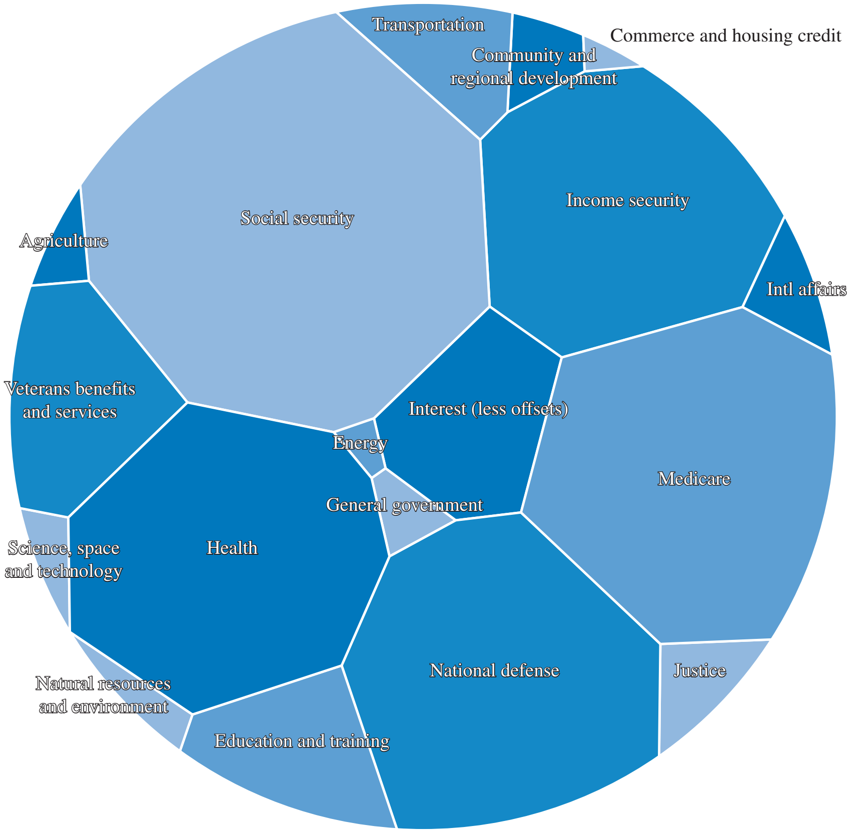
> Offensive content online is too often excused as not a big deal

and

> Many people take offensive content they see online too seriously.

Figure for Exercise 3.7

Where the federal government spends money, 2023



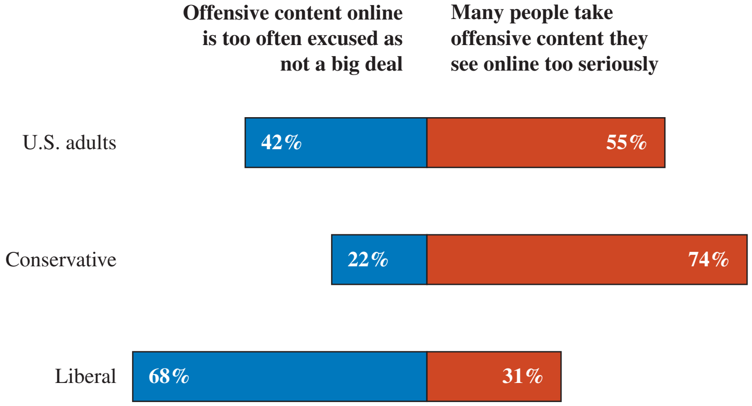
Source: Office of Management and Budget.

Copyright 2025 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s). Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

Figure for Exercise 3.8

About three-quarters of conservative Republicans say offensive content online is taken too seriously; about seven-in-ten liberal Democrats say it is too often excused

% of U.S. adults who say...

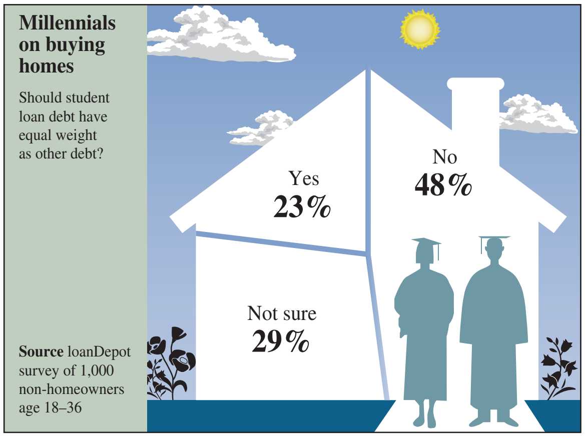


Note: Those who did not give an answer are not shown.
Source: For data on offensive content, survey of U.S. adults conducted September 8–13, 2020. For data on the appropriate balance of free speech and feeling safe online, survey of U.S. adults conducted July 13–19, 2020.

a. Explain how the sentence at the top of the display follows from the numbers in the graph.

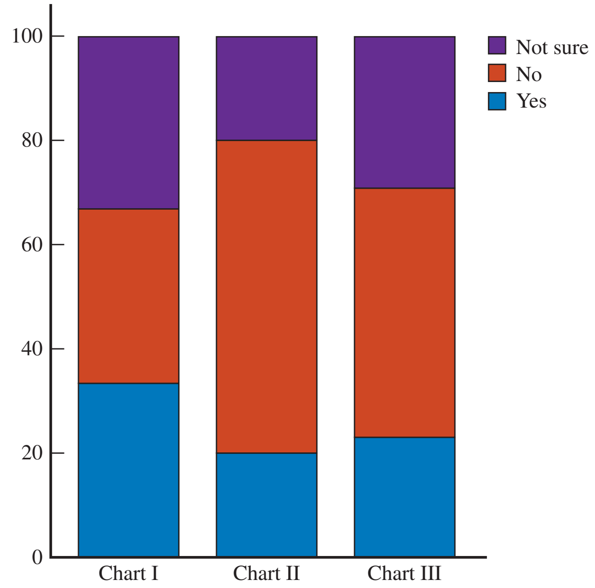
b. Write a few sentences summarizing the results of the survey and explaining how what you see in the graph does or does not support the title of the report.

3.9 The accompanying graphical display is similar to one that appeared in the ***USA TODAY Snapshots*** collection **(www.usatoday.com/)**. The graph is meant to represent responses to a question asking Millennials (people born between 1983 and 1996) whether they thought that student debt should be given equal weight with other debt when applying for a loan to buy a home.



a. Explain how this graph is similar to a pie chart.

b. Which of the following segmented bar charts (Chart I, Chart II, or Chart III) is a graph of the data used to create the graph that appeared in ***USA TODAY***? Explain.



3.10 ● Each year, *The Princeton Review* conducts surveys of high school students who are applying to college and of parents of college applicants. The report **"2020 College Hopes & Worries Survey Findings" (www.princetonreview.com)** included a summary of how 10,276 high school students responded to the question, "Ideally how far from home would you like the college you attend to be?" Students responded by choosing one of four possible distance categories. Also included was a summary of how 2,569 parents of students applying to college responded to the question, "How far from home would you like the college your child attends to be?" The accompanying frequency table summarizes the student and parent responses.

Copyright 2025 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s). Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

	Frequency	
Ideal Distance	**Students**	**Parents**
Less than 250 miles	3,494	1,233
250 to 500 miles	2,980	796
500 to 1,000 miles	2,055	334
More than 1,000 miles	1,747	206
Total	**10,276**	**2,569**

- **a.** Explain why you would want to use relative frequencies when constructing a comparative bar chart to compare ideal distance for students and parents.
- **b.** Construct a comparative bar chart for these data.
- **c.** Comment on similarities and differences in the distributions of ideal distance for parents and students.

3.11 The Gallup report **"Country Living Enjoys Renewed Appeal in the U.S." (gallup.com, January 5, 2021)** included data from polls of random samples of adult Americans conducted in 2018 and in 2020. People in the samples were asked the following question: If you could live anywhere you wished, would you live in a big city, a small city, a suburb of a big city, a suburb of a small city, a town, or a rural area?

Relative frequencies for the six response categories for the years 2018 and 2020 are given in the accompanying table.

Response	Relative Frequency for 2018	Relative Frequency for 2020
Big city	0.13	0.11
Small city	0.17	0.16
Suburb of a big city	0.21	0.16
Suburb of a small city	0.10	0.09
Town	0.12	0.17
Rural area	0.27	0.31

- **a.** Construct a comparative bar chart that shows the distributions of responses for 2018 and for 2020. (Hint: See Example 3.1. There should be two bars for each of the six response categories.)
- **b.** Comment on similarities and differences in the response distributions for 2018 and 2020.
- **c.** The title of the report was "Country Living Enjoys Renewed Appeal in U.S." How is this conclusion supported by the comparative bar chart?

3.12 During 2017, Gallup conducted a survey of adult Americans and asked the following question: What was the main reason you decided to enroll in the school or college where you completed your highest level of education? **("Why Higher Ed?," Gallup, Inc., January 2018)**. The responses are summarized in the accompanying table.

Location	28%
Access/Affordability	22%
School reputation and fit	20%
Good job or career	19%
Learning and knowledge	5%
Family or social expectations	4%
Other/No response	2%

- **a.** Construct a pie chart to summarize these data.
- **b.** Construct a bar chart to summarize these data.
- **c.** Which of these charts—a pie chart or a bar chart—best summarizes the important information? Explain.

3.13 The article **"Boomers Got Hooked on Stocks. Now They Can't Let Go." (*The Wall Street Journal*, June 18, 2023)** reported that in a large representative sample of people age 65 and over, 63% had money invested in the stock market and 37% did not have money invested in the stock market.

- **a.** Use the given information to construct a bar chart that uses relative frequency on the vertical axis.
- **b.** The article also included the following information: the percentage of people surveyed who had money invested in the stock market was 41% of those age 18 to 29, 67% for those age 30 to 49, and 66% for those age 50 to 64. Construct a comparative bar chart with bars for each of the two response categories for each age group. (Hint: See Figure 3.6.)

3.14 Heal the Bay is an environmental organization that releases an annual beach report card based on water quality **(Heal the Bay 2019–2020 Beach Report Card, www.beachreportcard.org)**. The grades for 19 beaches in San Luis Obispo County in California during wet winter weather were:

A A B A A A A A A A

A A A B A F A B A

- **a.** Summarize the San Luis Obispo wet winter weather grades by constructing a relative frequency distribution and a bar chart.
- **b.** The wet winter weather grades for 16 beaches in Santa Barbara County in California were:

  A A F B F F D A

  F F C F F A B F

  Construct a bar chart for the Santa Barbara wet winter weather grades.
- **c.** Do the bar charts from Parts (a) and (b) support the statement that beach wet weather water quality tends to be better in one county than the other? Explain.
- **d.** Construct a comparative bar chart for the grades for the two counties and comment on the differences between the two grade distributions.

Copyright 2025 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s). Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

Section 3.2 Displaying Numerical Data: Stem-and-Leaf Displays

A stem-and-leaf display is an effective and compact way to summarize numerical data. Each number in the data set is broken into two pieces, referred to as a stem and a leaf. The **stem** is the first part of the number and consists of the beginning digit(s). The **leaf** is the last part of the number and consists of the final digit(s). For example, the number 213 might be split into a stem of 2 and a leaf of 13 or a stem of 21 and a leaf of 3. The resulting stems and leaves are then used to construct the display.

Example 3.8 Dirty Money

Understand the context ➤

● Just how dirty is the money you carry in your wallet? The authors of the article **"Dirty Money: New Research Reveals the Filthiness of Our Cash, Cards, & Coins" (lendedu.com, 2020)** measured the bacteria on the surface of credit and debit cards, bills, and coins. Germ scores (with higher scores corresponding to dirtier surfaces) for 27 bills of various denominations ranging from $1 to $100 are given in the accompanying table. Leading zeros have been added to the germ scores that were less than 100 so that all the scores are represented as 3-digit numbers.

Consider the data ➤

032	035	039	041	054	055	056	068	074
078	098	104	105	108	119	119	126	129
143	163	200	242	259	273	349	609	633

Figure 3.14 shows a stem-and-leaf display for the germ score data. The numbers in the vertical column on the left of the display are the **stems**. Each number on the right of the vertical line is a **leaf** corresponding to one of the observations in the data set. The legend

Stem: Hundreds
Leaf: Ones

tells you that the observation that has a stem of 0 and a leaf of 32 corresponds to a germ score of 032 (0 hundreds and 32 ones). Similarly, the observation with the stem of 3 and leaf of 49 corresponds to a germ score of 349.

Figure 3.14
Stem-and-leaf display of germ scores.

0	32, 35, 39, 41, 54, 55, 56, 68, 74, 78, 98
1	04, 05, 08, 19, 19, 26, 29, 43, 63
2	00, 42, 59, 73
3	49
4	
5	
6	09, 33

Stem: Hundreds
Leaf: Ones

Interpret the results ➤

The display in Figure 3.14 suggests that a typical or representative germ score is in the stem 1 row, maybe around 119. The germ scores are mostly concentrated in the 32 to 273 range, but there are two values that stand out on the high end that are far removed from the rest of the data (609 and 633).

So what do these germ scores mean? For reference, it is recommended that a food establishment surface should have a germ score of 10 or less. The average germ scores for other surfaces studied were 68 for a New York City subway pole, 163 for bathrooms at Penn Station in New York City, and 664 for door handles at McDonald's!

The leaves on each line of the display in Figure 3.14 have been arranged in order from smallest to largest. Most statistical software packages order the leaves this way, but it is not necessary to do so to get an informative display that still shows many of the important characteristics of the data set, such as shape and variability.

● **Data set available online**

Stem-and-leaf displays can be useful to get a sense of a typical value for the data set, as well as a sense of how variable the values in the data set are. It is also easy to spot data values that are unusually far from the rest of the values in the data set. Such values are called outliers. In the stem-and-leaf display of the germ score data (Figure 3.14), the germ scores of 609 and 633 would be considered outliers.

Definition

Outlier: An unusually small or large data value. A precise rule for deciding when an observation is an outlier is given in Chapter 4.

Stem-and-Leaf Displays

When to Use Numerical data sets with a small to moderate number of observations (does not work well for very large data sets)

How to Construct

1. Select one or more leading digits for the stem values. The trailing digits (or sometimes just the first one of the trailing digits) become the leaves.
2. List possible stem values in a vertical column.
3. Record the leaf for every observation beside the corresponding stem value.
4. Indicate the units for stems and leaves someplace in the display.

What to Look For

- a representative or typical value in the data set
- the extent of variability about a typical value
- the presence of gaps and outliers in the data
- the extent of symmetry in the distribution of values
- the number and location of peaks

Example 3.9 Tuition at Public Universities

Understand the context ➤

● The introduction to this chapter gave data on average in-state tuition and fees at public institutions in the 2021–2022 academic year for the 50 U.S. states. The observations ranged from a low value of $3148 to a high value of $13,448. The data are reproduced here:

Consider the data ➤

7,861	7,324	5,037	4,554	3,148	7,106	8,741	7,643	3,795	4,737
5,020	6,259	6,155	10,304	6,709	5,282	6,496	6,772	8,732	8,403
7,142	8,964	7,238	7,551	5,004	4,731	3,542	6,547	5,904	10,084
10,103	3,653	5,395	7,546	7,152	6,285	7,702	12,190	10,338	8,601
7,971	6,825	6,476	6,085	8,537	13,448	5,995	6,308	6,424	4,429

Do the work ➤

Notice that some of the data values are only four digits (such as 7681) and others are five digits (such as 10,304). It is easiest to proceed if you first make all of the data values five digit numbers by adding leading zeros as shown:

07,861	07,324	05,037	04,554	03,148	07,106	08,741	07,643	03,795	04,737
05,020	06,259	06,155	10,304	06,709	05,282	06,496	06,772	08,732	08,403
07,142	08,964	07,238	07,551	05,004	04,731	03,542	06,547	05,904	10,084
10,103	03,653	05,395	07,546	07,152	06,285	07,702	12,190	10,338	08,601
07,971	06,825	06,476	06,085	08,537	13,448	05,995	06,308	06,424	04,429

A natural choice for the stem is the leading two digits. This would result in a display with 11 stems (03, 04, 05, 06, 07, 08, 09, 10, 11, 12, and 13). Using the first three digits of a number of the stem would result in 102 stems (031 to 134). A stem-and-leaf display with 102 stems would not be an effective summary of the data. In general, stem-and-leaf displays that use between 5 and 30 stems tend to work well.

● **Data set available online**

If you choose the first two digits as the stem, the remaining three digits (the hundreds, tens, and ones) would form the leaf. For example, for the first few values in the data set,

07,861 → stem = 07, leaf = 861
07,324 → stem = 07, leaf = 324
05,037 → stem = 05, leaf = 037

Interpret the results ➤

The leaves have been entered in the display of Figure 3.15 in the order they are encountered in the data set. Commas are used to separate the leaves only when each leaf has two or more digits. Figure 3.15 shows that most states had average in-state tuition and fees in the $5000 to $9000 range and that the typical average tuition and fees is around $7000. Six states have average in-state tuition and fees at public four-year institutions that are quite a bit higher than most other states (the six states with the highest values were Vermont, Pennsylvania, Rhode Island, Illinois, New Jersey, and New Hampshire).

Figure 3.15
Stem-and-leaf display of average tuition and fees.

03	148, 795, 542, 653
04	554, 737, 731, 429
05	037, 020, 282, 004, 904, 395, 995
06	259, 155, 709, 496, 772, 547, 285, 825, 476, 085, 308, 424
07	861, 324, 106, 643, 142, 238, 551, 546, 152, 702, 971
08	741, 732, 403, 964, 601, 537
09	
10	304, 084, 103, 338
11	
12	190
13	448

Stems: Thousands
Leaves: Ones

An alternative display (Figure 3.16) results from dropping all but the first digit of the leaf. This is what most statistical computer packages do when generating a display. Little information about typical value, variability, or shape is lost when this is done and the display is simpler and more compact.

Figure 3.16
Stem-and-leaf display of average in-state tuition and fees using truncated leaves.

03	1756
04	5774
05	0020939
06	217475284034
07	83161255179
08	774965
09	
10	3013
11	
12	1
13	1

Stems: Thousands
Leaves: Ones

Repeated Stems to Stretch a Display

Sometimes a natural choice of stems results in a display in which too many observations are concentrated on just a few stems. A more informative picture can be obtained by dividing the leaves at any given stem into two groups: those that begin with 0, 1, 2, 3, or 4 (the "low" leaves) and those that begin with 5, 6, 7, 8, or 9 (the "high" leaves). Then each stem value is listed twice when constructing the display, once for the low leaves and once again for the high leaves. It is also possible to repeat a stem more than twice. For example, each stem might be repeated five times, once for each of the leaf groupings {0, 1}, {2, 3}, {4, 5}, {6, 7}, and {8, 9}.

Example 3.10 Seat Belt Use

Understand the context ➤

● The accompanying data on seat belt use for each of the 50 U.S. states and the District of Columbia are from **"Traffic Safety Facts,"** which was published in June 2015 by the **National Highway Traffic Safety Administration**. The observations represent the percentage of drivers wearing seat belts in a large nationwide observational study.

Consider the data ➤

95.7	88.4	87.2	74.4	97.1	82.4	85.1	91.9	93.2	88.8	97.3	93.5	80.2
94.1	90.2	92.8	85.7	86.1	84.1	85.0	92.1	76.6	93.3	94.7	78.3	78.8
74.0	70.0	94.0	70.4	87.6	92.1	90.6	90.6	81.0	85.0	86.3	97.8	83.6
87.4	90.0	68.9	87.7	90.7	83.4	84.1	77.3	94.5	87.8	84.7	79.2	

The values in the data set range from 68.9% to 97.8%. Using the first two digits in each data value for the stem results in a large number of stems, while using only the first digit results in a stem-and-leaf display with only four stems.

Do the work ➤

The stem-and-leaf display using single-digit stems and leaves truncated to a single digit is shown in Figure 3.17. A stem-and-leaf display that uses repeated stems is shown in Figure 3.18. Here each stem is listed twice, once for the low leaves (those beginning with 0, 1, 2, 3, or 4) and once for the high leaves (those beginning with 5, 6, 7, 8, or 9). This display is more informative than the one in Figure 3.17, but it is much more compact than a display based on two-digit stems.

Figure 3.17
Stem-and-leaf display for the seat belt use data.

6	8
7	0 0 4 4 6 7 8 8 9
8	0 1 2 3 3 4 4 4 5 5 5 5 6 6 7 7 7 7 7 7 8 8
9	0 0 0 0 0 1 2 2 2 3 3 3 4 4 4 4 5 7 7 7

Figure 3.18
Stem-and-leaf display for the seat belt use data using repeated stems.

6H	8
7L	0 0 4 4
7H	6 7 8 8 9
8L	0 1 2 3 3 4 4 4
8H	5 5 5 5 6 6 7 7 7 7 7 7 8 8
9L	0 0 0 0 0 1 2 2 2 3 3 3 4 4 4 4
9H	5 7 7 7

Comparative Stem-and-Leaf Displays

Sometimes we want to see whether two data sets differ in some fundamental way. A comparative stem-and-leaf display, in which the leaves for one data set are listed to the right of the stem values and the leaves for the second data set are listed to the left, can provide a basis for making comparisons.

Example 3.11 Progress for Children

Understand the context ➤

● The report **"Progress for Children" (UNICEF, April 2005)** included the accompanying data on the percentage of primary-school-age children who were enrolled in school for 19 countries in Northern Africa and for 23 countries in Central Africa.

Consider the data ➤

Northern Africa

54.6	34.3	48.9	77.8	59.6	88.5	97.4	92.5	83.9	96.9	88.9
98.8	91.6	97.8	96.1	92.2	94.9	98.6	86.6			

Central Africa

58.3	34.6	35.5	45.4	38.6	63.8	53.9	61.9	69.9	43.0	85.0
63.4	58.4	61.9	40.9	73.9	34.8	74.4	97.4	61.0	66.7	79.6
98.9										

● **Data set available online**

Formulate a plan ➤

You can construct a comparative stem-and-leaf display using the first digit of each observation as the stem and the remaining two digits as the leaf. To keep the display simple the leaves can be truncated to one digit. For example, the observation 54.6 would be processed as

54.6 → stem = 5, leaf = 4 (truncated from 4.6)

and the observation 34.3 would be processed as

34.3 → stem = 3, leaf = 4 (truncated from 4.3)

The resulting comparative stem-and-leaf display is shown in Figure 3.19.

Figure 3.19
Comparative stem-and-leaf display for percentage of children enrolled in primary school.

Central Africa		Northern Africa
4854	3	4
035	4	8
838	5	49
6113913	6	
943	7	76
5	8	8386
87	9	7268176248

Stem: Tens
Leaf: Ones

What Do You Think?

How would you describe the differences in the distributions of percentage of children enrolled in primary school for the two regions?

Interpret the results ➤

From the comparative stem-and-leaf display you can see that there is quite a bit of variability in the percentage enrolled in school for both Northern and Central African countries and that the shapes of the two data distributions are quite different. The percentage enrolled in school tends to be higher in Northern African countries than in Central African countries, although the smallest value in each of the two data sets is about the same. For Northern African countries the distribution of values has a single peak in the 90s with the number of observations declining as you move toward the stems corresponding to lower percentages enrolled in school. For Central African countries the distribution is more symmetric, with a typical value in the mid 60s.

Exercises 3.15–3.23

● Data set available online

3.15 ● The **National Center for Health Statistics** provided the data in the accompanying table in the report **"National Vital Statistics Report" (cdc.gov, January 5, 2017)**. Entries in the table are the birth rates (births per 1000 of population) for the year 2015.

State	*Births	State	*Births
Alabama	12.3	Georgia	12.9
Alaska	15.3	Hawaii	12.9
Arizona	12.5	Idaho	13.8
Arkansas	13.1	Illinois	12.3
California	12.6	Indiana	12.7
Colorado	12.2	Iowa	12.6
Connecticut	10.0	Kansas	13.4
Delaware	11.8	Kentucky	12.6
District of Columbia	14.2	Louisiana	13.9
		Maine	9.5
Florida	11.1	Maryland	12.3

(continued)

State	*Births	State	*Births
Massachusetts	10.5	Oklahoma	13.6
Michigan	11.4	Oregon	11.8
Minnesota	12.7	Pennsylvania	11.0
Mississippi	12.8	Rhode Island	10.4
Missouri	12.3	South Carolina	11.9
Montana	12.2	South Dakota	14.4
Nebraska	14.1	Tennessee	12.4
Nevada	12.6	Texas	14.7
New Hampshire	9.3	Utah	16.9
New Jersey	11.5	Vermont	9.4
New Mexico	12.4	Virginia	12.3
New York	12.0	Washington	12.4
North Carolina	12.0	West Virginia	10.7
North Dakota	14.9	Wisconsin	11.6
Ohio	12.0	Wyoming	13.2

*Births per 1000 of population.

Construct a stem-and-leaf display using stems 9, 10, 11 . . . , 16. Comment on the interesting features of the display. (Hint: See Example 3.9.)

3.16 ● The paper **"State-Level Cancer Mortality Attributable to Cigarette Smoking in the United States" (*JAMA Internal Medicine* [2016]: 1792–1798)** included the following state estimates of the total number cancer deaths attributable to cigarette smoking in 2014.

State	Number of People	State	Number of People
Alabama	3,183	Montana	581
Alaska	296	Nebraska	927
Arizona	3,246	Nevada	1,535
Arkansas	2,175	New Hampshire	723
California	14,689	New Jersey	4,388
Colorado	1,876	New Mexico	964
Connecticut	1,774	New York	9,296
Delaware	591	North Carolina	5,844
District of Columbia	310	North Dakota	341
Florida	12,596	Ohio	7,598
Georgia	4,816	Oklahoma	2,441
Hawaii	642	Oregon	2,143
Idaho	731	Pennsylvania	7,931
Illinois	7,114	Rhode Island	631
Indiana	4,099	South Carolina	2,962
Iowa	1,793	South Dakota	476
Kansas	1,587	Tennessee	4,613
Kentucky	3,452	Texas	10,310
Louisiana	3,044	Utah	495
Maine	927	Vermont	382
Maryland	2,900	Virginia	4,110
Massachusetts	3,565	Washington	3,298
Michigan	6,232	West Virginia	1,581
Minnesota	2,552	Wisconsin	3,081
Mississippi	1,992	Wyoming	251
Missouri	4,047		

a. Construct a stem-and-leaf display using thousands as the stems and truncating the leaves to a single digit.

b. Write a few sentences describing the shape of the distribution and any unusual observations.

c. The four largest values were for California, Texas, Florida, and New York. Does this indicate that cancer deaths due to cigarette smoking is more of a problem in these states than elsewhere? Explain.

d. If you wanted to compare states on the basis of the cancer deaths due to cigarette smoking, would you use the data in the given table? If yes, explain why this would be reasonable. If no, what would you use instead as the basis for the comparison?

3.17 ● ***Business Insider* (businessinsider.com)** gave the following data on median age in 2021 for each of the 50 U.S. states and the District of Columbia (DC). Construct a stem-and-leaf display using stems 28, 29, … , 42. Comment on shape, center, and variability of the data distribution. Are there any unusual values in the data set that stand out?

State	Median Age	State	Median Age
Alabama	38.9	Montana	40.0
Alaska	34.5	Nebraska	36.5
Arizona	37.7	Nevada	38.0
Arkansas	38.1	New Hampshire	43.2
California	36.5	New Jersey	39.8
Colorado	36.8	New Mexico	37.7
Connecticut	40.9	New York	38.7
Delaware	40.1	North Carolina	38.8
DC	34.0	North Dakota	35.4
Florida	42.0	Ohio	39.3
Georgia	36.8	Oklahoma	36.6
Hawaii	39.2	Oregon	39.3
Idaho	36.3	Pennsylvania	40.8
Illinois	38.0	Rhode Island	39.5
Indiana	37.7	South Carolina	39.4
Iowa	38.3	South Dakota	36.9
Kansas	36.7	Tennessee	38.6
Kentucky	38.9	Texas	34.7
Louisiana	36.8	Utah	31.0
Maine	44.6	Vermont	42.6
Maryland	38.7	Virginia	38.2
Massachusetts	39.5	Washington	37.7
Michigan	39.8	West Virginia	42.4
Minnesota	37.9	Wisconsin	39.5
Mississippi	37.5	Wyoming	37.5
Missouri	38.5		

3.18 ● Example 3.10 gave data on seat belt use for each of the 50 U.S. states and the District of Columbia **("Traffic Safety Facts," National Highway Traffic Safety Administration, June 2015)**. The observations represent the percentage of drivers wearing seat belts in a large nationwide observational study.

Some, but not all, states enforce seat belts laws. Below are the seat belt usage data divided into two groups—states that enforce seat belts laws and those that do not.

States with Seat Belt Law Enforcement:

95.7 88.4 74.4 97.1 85.1 91.9 93.2 88.8 97.3 93.5 94.1 90.2
92.8 85.7 86.1 84.1 85.0 92.1 93.3 94.7 78.3 87.6 92.1 90.6
90.6 86.3 97.8 87.4 90.0 68.9 87.7 90.7 94.5 87.8 84.7

Copyright 2025 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s). Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

States without Seat Belt Law Enforcement:

87.2 82.4 80.2 76.6 78.8 74.0 70.0 94.0 70.4 81.0 85.0 83.6
68.9 83.4 84.1 77.3 79.2

a. Construct a comparative stem-and-leaf display using the tens digit as the stem and truncating the leaves to a single digit (the ones digit).

b. Write a few sentences commenting on similarities or differences in the seat belt use distributions for states with seat belt enforcement and states without seat belt enforcement.

3.19 ● The authors of the article **"Dirty Money: New Research Reveals the Filthiness of Our Cash, Cards, & Coins" (lendedu.com)** measured the bacteria on the surface of credit and debit cards, bills, and coins. Germ scores (with higher scores corresponding to dirtier surfaces) for 27 bills of various denominations ranging from $1 to $100 were given in Example 3.8 and are shown in the accompanying table. Leading zeros have been added to the germ scores that were less than 100 so that all the scores are represented as 3-digit numbers. The article also reported the germ scores for 12 coins, and those data are also given.

Germ Score for 27 Bills

032 035 039 041 054 055 056 068 074
078 098 104 105 108 119 119 126 129
143 163 200 242 259 273 349 609 633

Germ Score for 12 Coins

356 329 228 227 109 092 084 060 052
033 032 025

a. Construct a comparative stem-and-leaf display for the bills and coins.

b. Is the display constructed in Part (a) consistent with the following statement? Explain.
Statement: Bills tend to be dirtier than coins.

3.20 The article **"Volunteering Statistics for 2022: How Charitable Are We?" (teamstage.io/volunteering-statistics/)** included the following statement:

> Statistically, Utah is the most generous state in the U.S. The Beehive State tops the list of altruistic U.S. states, with 51% of the state's population reporting volunteer work. Volunteer statistics by the state show that Minnesota is the second most engaged state, with 45.1% of volunteers. Oregon, Alaska, and Iowa are also in the top five. Florida has the fewest volunteers, with just 22.8% of residents volunteering.

a. For the data set that includes the percentage who volunteered in 2022 for each of the 50 states, what is the largest value? What is the smallest value?

b. If you were going to construct a stem-and-leaf display for the data set consisting of the percentage who volunteered in 2022 for the 50 states, what stems would you use to construct the display? Explain your choice.

3.21 ● The accompanying table gives the U.S. gasoline tax per gallon data for each of the 50 states in 2022 (**forbes.com**, June 22, 2022)

State	Gasoline Tax (cents per gallon)
Alabama	28.0
Alaska	9.0
Arizona	18.0
Arkansas	24.5
California	51.1
Colorado	22.0
Connecticut	25.0
Delaware	23.0
Florida	19.0
Georgia	29.1
Hawaii	16.0
Idaho	32.0
Illinois	39.2
Indiana	32.0
Iowa	30.5
Kansas	24.0
Kentucky	24.6
Louisiana	20.0
Maine	30.0
Maryland	36.1
Massachusetts	24.0
Michigan	27.2
Minnesota	28.5
Mississippi	18.0
Missouri	19.5
Montana	32.5
Nebraska	24.8
Nevada	23.0
New Hampshire	22.2
New Jersey	42.1
New Mexico	17.0
New York	25.4
North Carolina	38.5
North Dakota	23.0
Ohio	38.5
Oklahoma	19.0
Oregon	38.0
Pennsylvania	57.6
Rhode Island	34.0
South Carolina	26.0
South Dakota	28.0
Tennessee	26.0
Texas	20.0
Utah	31.9
Vermont	31.6
Virginia	26.2
Washington	49.4
West Virginia	35.7
Wisconsin	30.9
Wyoming	24.0

- **a.** Construct a stem-and-leaf display of these data.
- **b.** Based on the stem-and-leaf display, what do you notice about the center and spread of the data distribution?
- **c.** Do any values in the data set stand out as unusual? If so, which states correspond to the unusual observations, and how do these values differ from the rest?

3.22 ● The *World Population Review* gives the following data on percentage population growth in 2021 for the 50 U.S. states **(worldpopulationreview.com)**. Each state is also classified as belonging to the eastern or western part of the United States.

State	2021 Growth (%)	East/West	State	2021 Growth (%)	East/West
Alabama	1.0	E	Montana	2.3	W
Alaska	−1.5	W	Nebraska	1.4	W
Arizona	5.1	W	Nevada	5.2	W
Arkansas	0.8	E	New Hampshire	1.4	E
California	0.4	E	New Jersey	−0.1	E
Colorado	3.6	E	New Mexico	0.6	W
Connecticut	−0.5	W	New York	−1.2	E
Delaware	2.6	W	North Carolina	3.1	E
Florida	3.3	E	North Dakota	1.6	W
Georgia	3.0	E	Ohio	0.3	E
Hawaii	−1.0	W	Oklahoma	1.3	W
Idaho	6.3	W	Oregon	2.6	W
Illinois	−1.2	E	Pennsylvania	0.3	E
Indiana	1.7	E	Rhode Island	0.3	E
Iowa	0.6	E	South Carolina	3.8	E
Kansas	0.2	W	South Dakota	2.0	W
Kentucky	0.4	E	Tennessee	2.6	E
Louisiana	−0.7	E	Texas	3.9	W
Maine	1.2	E	Utah	5.0	W
Maryland	0.5	E	Vermont	−0.2	E
Massachusetts	0.4	E	Virginia	1.2	E
Michigan	0.1	E	Washington	3.6	W
Minnesota	1.8	E	West Virginia	−2.0	E
Mississippi	−0.5	E	Wisconsin	0.8	E
Missouri	0.8	E	Wyoming	0.6	W

- **a.** Construct a stem-and-leaf display for the entire data set. Hint: Because some of the data values are negative, you will need to use stems of −2, −1, −0 (for the negative values that have a first digit of 0), 0 (for the positive values that have a first digit of 0), 1, 2, 3, 4, 5, and 6.
- **b.** Comment on any interesting features of the display. Do any of the observations appear to be outliers?
- **c.** Now construct a comparative stem-and-leaf display for the Eastern and Western states. Write a few sentences comparing the two distributions.

3.23 ● The percentage of teens not in school or working in 2020 for the 50 states were given in the **2022 Kids Count Data Book (aecf.org)** and are shown in the following table:

State	Percentage	State	Percentage
Alabama	8	Montana	8
Alaska	10	Nebraska	4
Arizona	8	Nevada	10
Arkansas	9	New Hampshire	5
California	7	New Jersey	5
Colorado	6	New Mexico	10
Connecticut	5	New York	6
Delaware	6	North Carolina	7
Florida	7	North Dakota	5
Georgia	8	Ohio	6
Hawaii	9	Oklahoma	8
Idaho	7	Oregon	8
Illinois	6	Pennsylvania	6
Indiana	6	Rhode Island	4
Iowa	6	South Carolina	8
Kansas	6	South Dakota	6
Kentucky	9	Tennessee	7
Louisiana	10	Texas	8
Maine	6	Utah	6
Maryland	6	Vermont	5
Massachusetts	5	Virginia	5
Michigan	7	Washington	7
Minnesota	4	West Virginia	9
Mississippi	9	Wisconsin	5
Missouri	7	Wyoming	6

Note that the percentages range from a low of 4% to a high of 10%. In constructing a stem-and-leaf display for these data, if you regard each percentage as a two-digit number and use the first digit for the stem, then there are only two possible stems, 0 and 1. One solution is to use repeated stems. Consider a scheme that divides the leaf range into five parts: 0 and 1, 2 and 3, 4 and 5, 6 and 7, and 8 and 9. Then, for example, stem 0 could be repeated as

0	with leaves 0 and 1
0t	with leaves 2 and 3
0f	with leaves 4 and 5
0s	with leaves 6 and 7
0*	with leaves 8 and 9

Construct a stem-and-leaf display for this data set that uses stems 0f, 0s, 0*, and 1, 1t, and 1f. Comment on the important features of the display. (Hint: See Example 3.10.)

Copyright 2025 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s). Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

Section 3.3 Displaying Numerical Data: Frequency Distributions and Histograms

A stem-and-leaf display is not always an effective way to summarize data because it can be unwieldy when the data set contains a large number of observations. Frequency distributions and histograms are displays that work well for large data sets.

Frequency Distributions and Histograms for Discrete Numerical Data

Discrete numerical data often results from counting. When this is the case, each observation is a whole number. As in the case of categorical data, a frequency distribution for discrete numerical data lists each possible value (either individually or grouped into intervals), the associated frequency, and sometimes the corresponding relative frequency. Recall that a relative frequency is calculated by dividing a frequency by the total number of observations in the data set.

Example 3.12 Promiscuous Queen Bees

Understand the context ➤

● Queen honey bees mate shortly after they become adults. During a mating flight, the queen usually takes multiple partners, collecting sperm that she will store and use throughout the rest of her life. The authors of the paper **"The Curious Promiscuity of Queen Honey Bees" (*Annals of Zoology* [2001]: 255–265)** studied the behavior of 30 queen honey bees to learn about the length of mating flights and the number of partners a queen takes during a mating flight.

The accompanying data on number of partners on one mating flight were generated to be consistent with summary values and graphs given in the paper.

Consider the data ➤

Number of Partners

12	2	4	6	6	7	8	7	8	11
8	3	5	6	7	10	1	9	7	6
9	7	5	4	7	4	6	7	8	10

The corresponding relative frequency distribution is given in Table 3.1. The smallest value in the data set is 1 and the largest is 12, so the possible values from 1 to 12 are listed in the table, along with the corresponding frequency and relative frequency.

Table 3.1 Relative Frequency Distribution for Number of Partners

Number of Partners	Frequency	Relative Frequency	
1	1	0.033	← $\frac{1}{30} = 0.033$
2	1	0.033	
3	1	0.033	
4	3	0.100	
5	2	0.067	
6	5	0.167	
7	7	0.233	
8	4	0.133	
9	2	0.067	
10	2	0.067	
11	1	0.033	
12	1	0.033	
Total	**30**	**0.999**	← *Differs from 1 due to rounding*

● **Data set available online**

Copyright 2025 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s). Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

Interpret the results ➤

From the relative frequency distribution, you can see that five of the queen bees had six partners during their mating flight. The corresponding relative frequency, $\frac{5}{30} = 0.167$, tells you that the proportion of queens with six partners is 0.167, or equivalently 16.7% of the queens had six partners. Adding the relative frequencies for the values 10, 11, and 12 gives

$$0.067 + 0.033 + 0.033 = 0.133$$

indicating that 13.3% of the queens had 10 or more partners.

It is possible to create a more compact frequency distribution by grouping some of the possible values into intervals. For example, you might group together 1, 2, and 3 partners to form an interval of 1–3, with a corresponding frequency of 3. The grouping of other values in a similar way results in the relative frequency distribution shown in Table 3.2.

Table 3.2 **Relative Frequency Distribution of Number of Partners Using Intervals**

Number of Partners	Frequency	Relative Frequency
1–3	3	0.100
4–6	10	0.333
7–9	13	0.433
10–12	4	0.133

A histogram for discrete numerical data is a graph of the frequency or relative frequency distribution, and it is similar to the bar chart for categorical data. Each frequency or relative frequency is represented by a rectangle centered over the corresponding value (or range of values) and the area of the rectangle is proportional to the corresponding frequency or relative frequency.

Histogram for Discrete Numerical Data

When to Use Discrete numerical data. Works well for large data sets.

How to Construct

1. Draw a horizontal scale, and mark the possible values of the variable.
2. Draw a vertical scale, and mark it with either frequency or relative frequency.
3. Above each possible value, draw a rectangle centered at that value (so that the rectangle for 1 is centered at 1, the rectangle for 5 is centered at 5, and so on). The height of each rectangle is determined by the corresponding frequency or relative frequency. Often possible values are consecutive whole numbers, in which case the base width for each rectangle is 1.

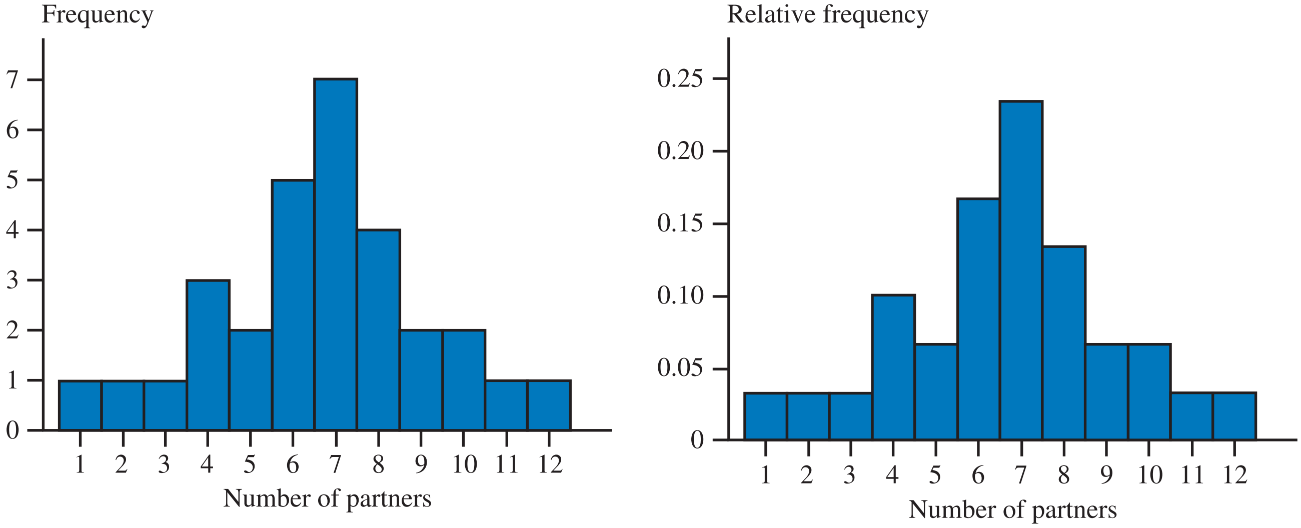
What to Look For

- Center or typical value
- Extent of variability
- General shape
- Location and number of peaks
- Presence of gaps and outliers

Example 3.13 Promiscuous Queen Bees Revisited

The queen bee data of Example 3.12 were summarized in a frequency distribution. The corresponding histogram is shown in Figure 3.20. Note that each rectangle in the histogram is centered over the corresponding value. When relative frequency instead of frequency is used for the vertical scale, the scale on the vertical axis is different but all essential characteristics of the graph (shape, center, variability) are unchanged.

Figure 3.20
Histogram and relative frequency histogram of queen bee data.



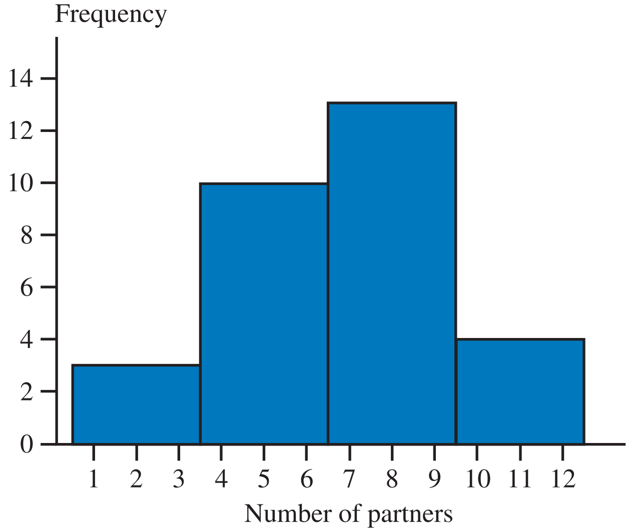
What Do You Think?

Based on the histograms in Figure 3.20, what would you say is a typical value for number of partners on a mating flight?

Are the values in the data set similar or is there quite a bit of variability? How is this reflected in the histogram?

A histogram based on the grouped frequency distribution of Table 3.2 can be constructed in a similar fashion, and is shown in Figure 3.21. A rectangle represents the frequency or relative frequency for each interval. For the interval from 1 to 3, the rectangle extends from 0.5 to 3.5 so that there are no gaps between the rectangles of the histogram.

Figure 3.21
Histogram of queen bee data using intervals.



Sometimes a discrete numerical data set contains a large number of possible values and perhaps also has a few large or small values that are far away from most of the data. In this case, rather than forming a frequency distribution with a very long list of possible values, it is common to group the observed values into intervals or ranges. This is illustrated in Example 3.14.

Copyright 2025 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s). Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

Example 3.14 Identity Theft

Understand the context ➤

● Identity theft occurs when someone steals your personal information to commit fraud, and it is a growing problem as more people complete more transactions online. The data in the accompanying table are the number of identity theft reports per 100,000 population for the 50 U.S. states from the report **"Identity Theft and Credit Card Fraud Statistics for 2020" (The Motley Fool, 2021)**.

Consider the data ➤

State	Identity Theft Reports per 100,000Population	State	Identity Theft Reports per 100,000Population
Alabama	173	Montana	67
Alaska	73	Nebraska	68
Arizona	150	Nevada	256
Arkansas	150	New Hampshire	96
California	257	New Jersey	205
Colorado	110	New Mexico	100
Connecticut	128	New York	186
Delaware	226	North Carolina	179
Florida	304	North Dakota	59
Georgia	427	Ohio	118
Hawaii	95	Oklahoma	94
Idaho	81	Oregon	96
Illinois	182	Pennsylvania	163
Indiana	95	Rhode Island	108
Iowa	61	South Carolina	213
Kansas	78	South Dakota	47
Kentucky	67	Tennessee	158
Louisiana	227	Texas	256
Maine	60	Utah	149
Maryland	210	Vermont	54
Massachusetts	125	Virginia	121
Michigan	135	Washington	94
Minnesota	80	West Virginia	59
Mississippi	158	Wisconsin	86
Missouri	121	Wyoming	55

What Do You Think?

Why do you think that the report used identity theft reports per 100,000 people rather than just the total number of identity theft reports for each state?

A relative frequency distribution is given in Table 3.3 and the corresponding histogram is shown in Figure 3.22. Notice that the possible values have been grouped into intervals rather than listing each of the possible values, which ranged from 47 to 427. This results in a much more compact frequency distribution and histogram that still communicates the important features of the data set.

Interpret the results ➤

From the relative frequency distribution and histogram, you can see that while there is a lot of variability in the number of identity theft reports per 100,000 people for the 50 states, the majority were in the 50 to 200 range and a typical number is in the 100 to 150 range. There is one state (Georgia) that stands out as unusual, with 427 identity theft reports per 100,000 people.

Also note that the rectangle representing the interval 50 to 99 actually extends from 49.5 to 99.5. This is similar to what happens in histograms for discrete numerical data where

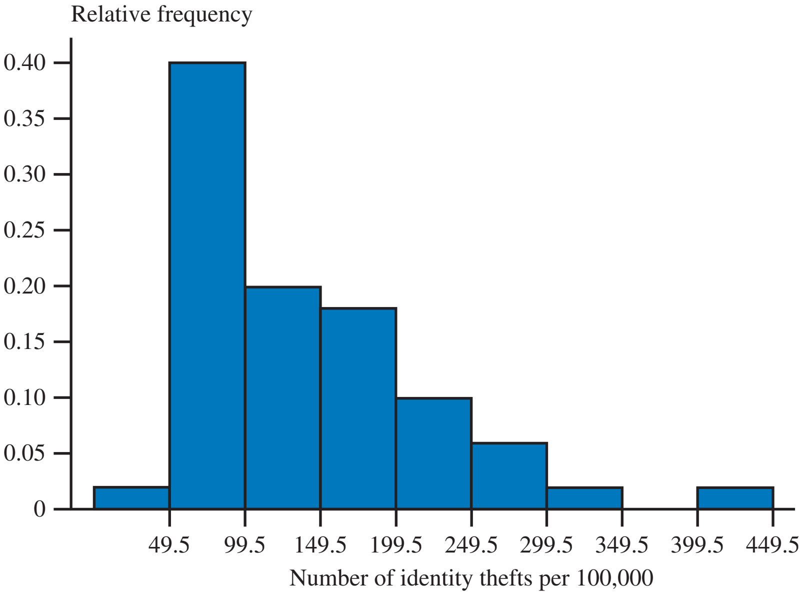
● **Data set available online**

there is no grouping. For example, in Figure 3.20 the rectangle representing 2 is centered at 2 but extends from 1.5 to 2.5 on the number of partners scale.

Table 3.3 **Relative Frequency Distribution for the Identity Theft Data**

Number of Identity Thefts per 100,000	Frequency	Relative Frequency
0–49	1	0.02
50–99	20	0.40
100–149	10	0.20
150–199	9	0.18
200–249	5	0.10
250–299	3	0.06
300–349	1	0.02
350–399	0	0.00
400–449	1	0.02

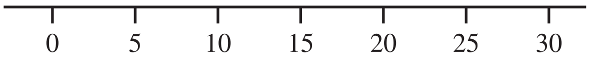
Figure 3.22
Relative frequency histogram for the identity theft data.



Frequency Distributions and Histograms for Continuous Numerical Data

For continuous data, such as observations on reaction time (in seconds) or weight of airline passenger carry-on luggage (in pounds), there are no natural categories. In this case, you can define your own categories. For carry-on luggage weight, you might expect weights up to about 30 pounds. One way to group the weights into 5-pound intervals is shown in Figure 3.23. Then each observed data value could be classified into one of these intervals. The intervals used are sometimes called **class intervals**. The class intervals play the same role that the categories or individual values played in frequency distributions for categorical or discrete numerical data.

Figure 3.23
Suitable class intervals for carry-on luggage weight data.



There is one further difficulty that needs to be addressed. Where should you place an observation such as 20, which falls on a boundary between class intervals? One convention is to define intervals so that such an observation is placed in the upper rather than the lower class interval. This means that in a frequency distribution, one class might be 15 to < 20, where the symbol < is a substitute for the phrase *less than*. This class interval will contain all observations that are greater than or equal to 15 and less than 20. The observation 20 would then fall in the class interval 20 to < 25.

Example 3.15 The High Cost of Housing

Understand the context ➤

● Many people struggle to afford rental housing, and the cost of rental housing varies widely from state to state. The report **"Out of Reach" (National Low Income Housing Coalition, 2021, nlihc.org)** states that 30 states now have minimum wages higher than the federal minimum wage. But even taking the higher minimum wage into account, for the 50 states the average number of hours per week needed to afford a two-bedroom rental home was 91.40 hours. The report includes estimates of the hourly wage a full-time worker would need to earn in order to afford a typical two-bedroom rental home without spending more than 30% of their income on housing for each of the 50 U.S. states and the District of Columbia.

Consider the data ➤

29.31 25.14 39.03 23.74 37.69 21.83 17.36 20.21 22.3 17.76 16.85 27.5 17.22
21.98 16.28 16.81 16.91 15.46 16.61 21.78 15.97 16.66 14.6 17.82 15.21 15.8
17.59 15.78 22.11 16.57 17.89 18.55 16.64 14.83 24.82 19.42 18.08 18.46 24.41
19.95 34.03 21.39 27.37 33.94 22.52 36.24 28.6 31.96 24.73 22.54 23.68

The smallest observation is \$14.60 (for Arkansas). It is reasonable to start the first class interval at \$10 and let each interval have a width of 5. This results in class intervals starting with \$10 to < \$15 and continuing up to \$35 to < \$40 to accommodate the largest value of \$39.03 (for California).

Table 3.4 displays the resulting frequency distribution, along with the relative frequencies.

Table 3.4 Frequency Distribution for Hourly Wage Needed for a Full-Time Worker to Afford a Two-Bedroom Rental

Class Interval	Frequency	Relative Frequency
\$10 to < \$15	2	0.04
\$15 to < \$20	24	0.47
\$20 to < \$25	14	0.27
\$25 to < \$30	5	0.10
\$30 to < \$35	3	0.06
\$35 to < \$40	3	0.06
Total	**51**	**1.00**

Interpret the results ➤

Various relative frequencies can be combined to yield other interesting information. For example,

$$\begin{pmatrix}\text{proportion of states where}\\ \text{the hourly wage would}\\ \text{need to be more than \$30}\end{pmatrix} = \begin{pmatrix}\text{proportion in}\\ \text{\$30 to} < \text{\$35 interval}\end{pmatrix} + \begin{pmatrix}\text{proportion in}\\ \text{\$35 to} < \text{\$40 interval}\end{pmatrix}$$

$$= 0.06 + 0.06 = 0.12\ (12\%)$$

● **Data set available online**

and

$$\begin{pmatrix}\text{proportion of states}\\ \text{where the hourly wage}\\ \text{would need to be}\\ \text{between \$25 and \$40}\end{pmatrix} = \begin{pmatrix}\text{proportion in}\\ \text{\$25 to} < \text{\$30}\\ \text{interval}\end{pmatrix} + \begin{pmatrix}\text{proportion in}\\ \text{\$30 to} < \text{\$35}\\ \text{interval}\end{pmatrix} + \begin{pmatrix}\text{proportion in}\\ \text{\$35 to} < \text{\$40}\\ \text{interval}\end{pmatrix}$$

$$= 0.27 + 0.10 + 0.06 = 0.43\ (43\%)$$

There are no set rules for selecting either the number of class intervals or the length of the intervals. Using a few relatively wide intervals will bunch the data, whereas using a great many relatively narrow intervals may spread the data over too many intervals, so that no interval contains more than a few observations. Neither type of distribution will give an informative picture of how data values are distributed, and interesting features of the data set may be missed. In general, with a small number of observations, relatively few intervals, perhaps between 5 and 10, should be used. With a large number of observations, a distribution based on 15 to 20 (or even more) intervals is often recommended. The quantity

$$\sqrt{\text{number of observations}}$$

is sometimes used as an estimate of an appropriate number of intervals: 5 intervals for 25 observations, 10 intervals when the number of observations is 100, and so on.

Histograms for Continuous Numerical Data

When the class intervals in a frequency distribution are all of equal width, it is easy to construct a histogram using the information in a frequency distribution.

Histogram for Continuous Numerical Data When the Class Interval Widths Are Equal

When to Use Continuous numerical data. Works well, even for large data sets.

How to Construct

1. Mark the boundaries of the class intervals on a horizontal axis.
2. Use either frequency or relative frequency on the vertical axis.
3. Draw a rectangle for each interval directly above the corresponding interval (so that the edges are at the class interval boundaries). The height of each rectangle is the corresponding frequency or relative frequency.

What to Look For

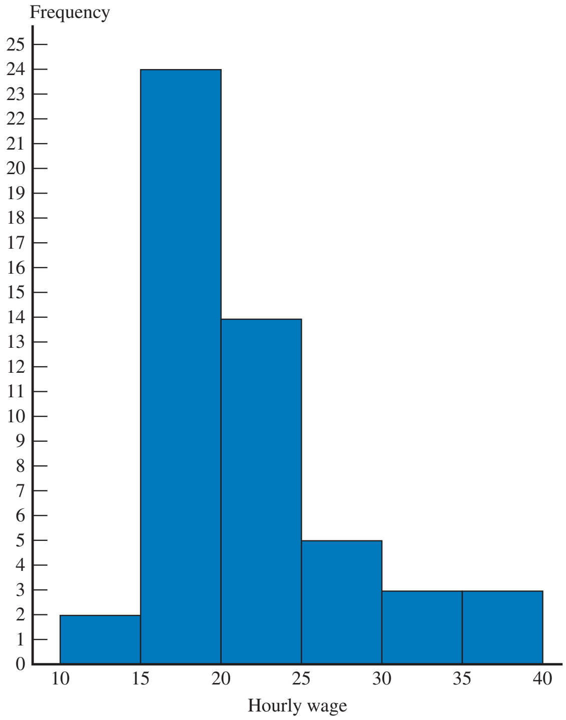
- Center or typical value
- Extent of variability
- General shape
- Location and number of peaks
- Presence of gaps and outliers

Two people making reasonable and similar choices for the number of intervals, the interval width, and the starting point of the first interval, will usually obtain histograms that are similar in terms of shape, center, and variability.

Example 3.16 The High Cost of Housing

The data on hourly wage needed in order for a full-time worker to afford a two-bedroom rental given in Example 3.15 were used to create a histogram using Minitab (Figure 3.24). Based on the histogram, you can see that an hourly wage above $20 would be needed to afford a two-bedroom rental in about half of the states and that the hourly wage needed is over $30 for quite a few states.

Figure 3.24
Histogram of the hourly wage data.



Example 3.17 Sleep Deficit and School Start Time

Understand the context ➤

The authors of the paper **"The Influence of School Time on Sleep Patterns of Children and Adolescents" (*Sleep Medicine* [2016]: 33–39)** were interested in determining if early school start time has an effect on the amount of sleep school-age children get. They studied students in Brazil in a region that offered both a morning school start time (with classes from 7:30 a.m. to noon) and an afternoon school start time (with classes from 1:30 p.m. to 5:30 p.m.). One variable of interest in the study was sleep deficit, which they defined as the difference in sleep duration on weekends and the sleep duration on nights with school the next day. Sleep deficit was measured in hours, so a student that typically slept 9 hours a night on weekends and only 7 hours a night on weeknights would have a sleep deficit of 2 hours. A student with a sleep deficit of −3 hours would be one who typically slept 3 hours longer on weeknights than on weekends, resulting in a negative difference.

Table 3.5 gives relative frequencies (as approximate values based on a graph that appears in the paper) for various sleep deficit intervals for students with the morning start time and for students with the afternoon start time. Relative frequencies are used because there were 538 students in the morning start group and only 101 students in the afternoon start group.

Table 3.5 Relative Frequency Distribution of Sleep Deficit

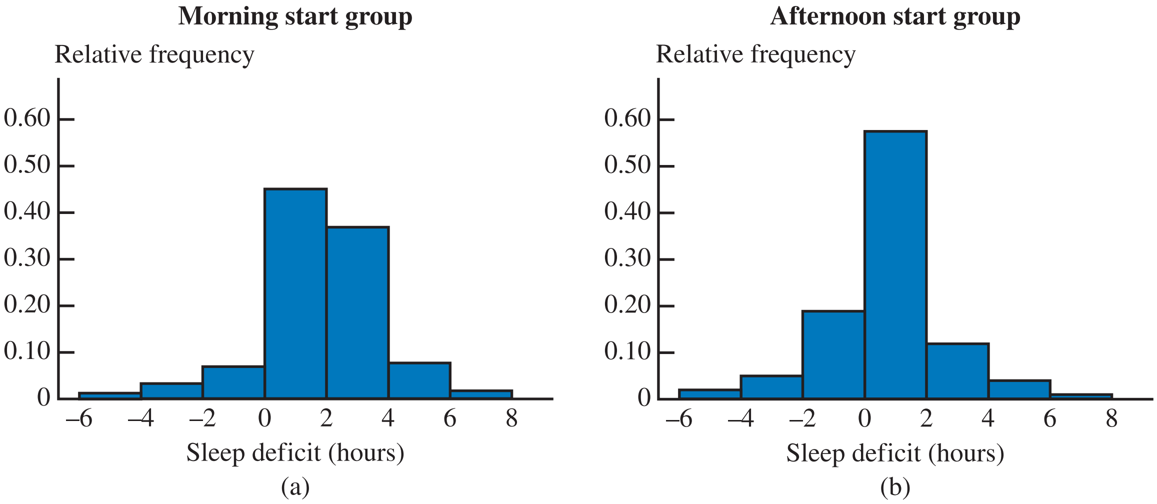
Consider the data ➤

Sleep Deficit (in hours)	Morning Start Relative Frequency	Afternoon Start Relative Frequency
-6 to < -4	0.007	0.020
-4 to < -2	0.028	0.050
-2 to < 0	0.065	0.190
0 to < 2	0.442	0.570
2 to < 4	0.364	0.120
4 to < 6	0.078	0.040
6 to < 8	0.015	0.010

Figure 3.25(a) is the relative frequency histogram for the morning start time group and Figure 3.25(b) is the relative frequency histogram for afternoon start time group. To be able to compare the distributions of sleep deficit for the morning start and afternoon start groups, it is important to use the same scales for the two histograms.

Figure 3.25

Histogram of sleep deficit: (a) morning start; (b) afternoon start.



Interpret the results ➤

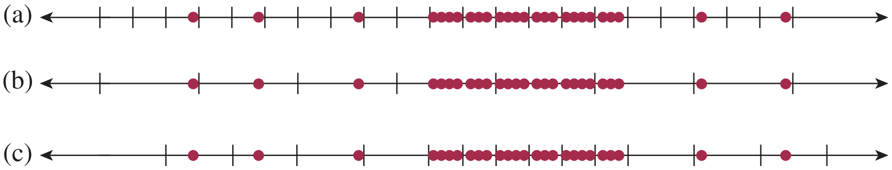
Notice that both histograms have a single peak, with the majority of students in both groups having positive values for sleep deficit. However, students in the afternoon school start time group tended to have smaller deficits than the students in the morning start time group, indicating that difference between sleep duration on weekends and sleep duration on weekdays tended to be smaller for students with an afternoon start time. This is one reason that many people are recommending later school start times.

Class Intervals of Unequal Widths

Figure 3.26 shows a data set in which a large number of observations are concentrated in the middle of the data set, with only a few unusual values far below and far above the center of the data. If a frequency distribution is based on short intervals of equal width, many intervals will be required to capture all observations, and a lot of them will contain no observations, as shown in Figure 3.26(a). On the other hand, only a few wide intervals will capture all values, but then most of the observations will be grouped into just a few intervals, as shown in Figure 3.26(b). In such situations, it is best to use a combination of wide class intervals where there are few data points and shorter intervals where there are many data points, as shown in Figure 3.26(c).

Figure 3.26

Three choices of class intervals for a data set with outliers: (a) many short intervals of equal width; (b) a few wide intervals of equal width; (c) intervals of unequal width.



Constructing a Histogram for Continuous Data When Class Interval Widths Are Unequal

When class intervals are not all the same width, frequencies or relative frequencies should not be used on the vertical axis of a histogram. Instead, the height of each rectangle, called the **density** for the class interval, is given by

$$\text{density} = \text{rectangle height} = \frac{\text{relative frequency of class interval}}{\text{class interval width}}$$

The vertical axis is called the **density scale**.

Copyright 2025 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s). Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

The use of the density scale to construct the histogram ensures that the area of each rectangle in the histogram will be proportional to the corresponding relative frequency. The formula for density can also be used when class widths are equal. However, when the intervals all have equal width, the extra arithmetic required to obtain the densities is not necessary.

Example 3.18 Student Debt on the Rise

Understand the context ➤

At many U.S. colleges and universities, low-income students often need to take on a large debt burden to pay for their education. The article **"Poor Feel the Bite of Rising College Costs" (*The Wall Street Journal*, February 20, 2016)** classified 1319 colleges into intervals based on the median yearly debt for students from families with an annual income of $30,000 or less. The data are summarized in the frequency distribution given in Table 3.6.

Table 3.6 Frequency Distribution for Median Student Debt

Consider the data ➤

Class Interval	Frequency	Relative Frequency	Interval Width	Density
\$0 to < \$10,000	166	0.126	\$10,000	0.000013
\$10,000 to < \$15,000	389	0.295	\$5,000	0.000059
\$15,000 to < \$20,000	456	0.346	\$5,000	0.000069
\$20,000 to < \$40,000	308	0.234	\$20,000	0.000012

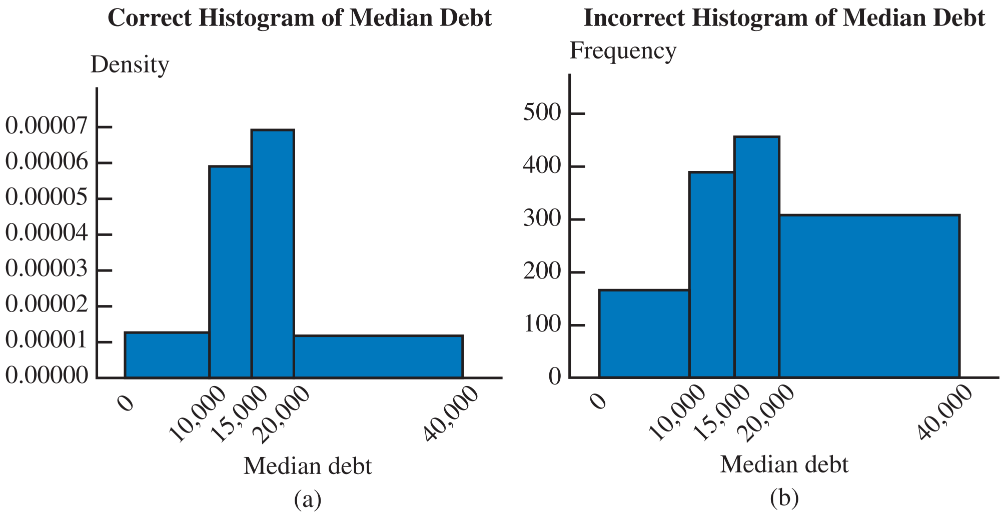
Looking at the frequency distribution, you can see that the median student debt for low-income students was between \$20,000 and \$40,000 for 308 of the colleges and was between \$15,000 and \$20,000 for 456 of the colleges. Notice also that the income intervals used in the article are not of equal width. Two intervals have a width of \$5000 (for example, the interval from \$10,000 to < \$15,000), and the other two intervals have widths of \$10,000 and \$20,000, respectively.

Figure 3.27 displays two histograms based on this frequency distribution. The histogram in Figure 3.27(a) is correctly drawn, with density used to determine the height of each bar. The histogram in Figure 3.27(b) has height equal to relative frequency and is therefore not correct.

What Do You Think?

Why is the histogram in Figure 3.27(b) misleading?

Notice that the histogram in Figure 3.27(b) exaggerates the proportion of colleges with low median debt and the proportion of colleges with very high median debt—the areas of the two most extreme rectangles are much too large. The eye is naturally drawn to large areas, so it is important that the areas correctly represent the relative frequencies.



(a)

(b)

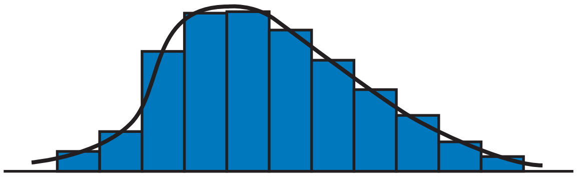
Figure 3.27
Histograms for median student debt:
(a) a correct histogram (height = density);
(b) an incorrect histogram (height = relative frequency).

Copyright 2025 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s). Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

Histogram Shapes

General shape is an important characteristic of a histogram. In describing various shapes it is convenient to approximate the histogram with a smooth curve (called a *smoothed histogram*). This is illustrated in Figure 3.28.

Figure 3.28
Approximating a histogram with a smooth curve.



One description of general shape relates to the number of peaks, or **modes**.

Definitions

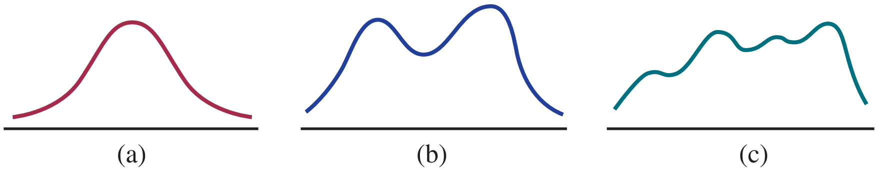
Unimodal: A histogram is **unimodal** if it has a single peak.

Bimodal: A histogram is **bimodal** if it has two peaks.

Multimodal: A histogram is **multimodal** if it has more than two peaks.

These shapes are illustrated in Figure 3.29.

Figure 3.29
Smoothed histograms with various numbers of modes: (a) unimodal; (b) bimodal; (c) multimodal.

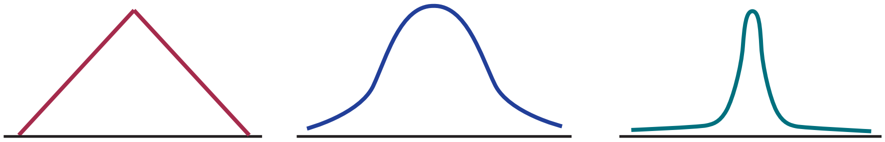


Bimodal histograms sometimes occur when the data set consists of observations on two quite different kinds of individuals or objects. For example, consider a large data set consisting of driving times for automobiles traveling between San Luis Obispo, California, and Monterey, California. This histogram would show two peaks, one for those cars that took the inland route (roughly 2.5 hours) and another for those cars traveling up the coast highway (3.5–4 hours).

However, bimodality does not automatically follow in such situations. Bimodality will occur in the histogram of the combined groups only if the centers of the two separate histograms are far apart relative to the variability in the two data sets. For example, a large data set consisting of heights of college students would probably not produce a bimodal histogram because the typical height for males (about 69 in.) and the typical height for females (about 66 in.) are not very far apart relative to the variability in the height distributions.

Unimodal histograms come in a variety of shapes. A unimodal histogram is **symmetric** if there is a vertical line of symmetry such that the part of the histogram to the left of the line is a mirror image of the part to the right. (Bimodal and multimodal histograms can also be symmetric in this way.) Several different symmetric smoothed histograms are shown in Figure 3.30.

Figure 3.30
Several symmetric unimodal smoothed histograms.



Copyright 2025 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s). Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

Proceeding to the right from the peak of a unimodal histogram, you move into what is called the **upper tail** of the histogram. Going in the opposite direction moves you into the **lower tail**.

Definitions

Skewed: A unimodal histogram that is not symmetric is described as **skewed**.

Positively skewed or right skewed: A skewed histogram in which the upper tail of the histogram stretches out much farther than the lower tail is described as **positively skewed** or **right skewed**.

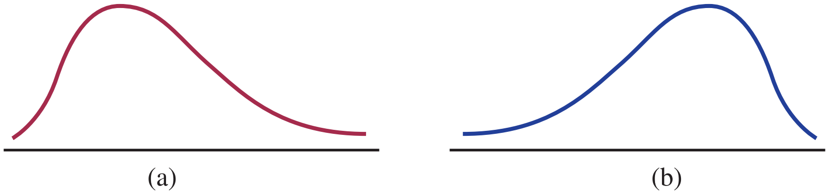
Negatively skewed or left skewed: A skewed histogram in which the lower tail of the histogram stretches out much farther than the upper tail is described as **negatively skewed** or **left skewed**.

These two types of skewness are illustrated in Figure 3.31. Positive skewness is much more frequently encountered than is negative skewness. An example of positive skewness occurs in the distribution of single-family home prices in Los Angeles County. Most homes are moderately priced (at least for California), whereas the relatively few homes in Beverly Hills and Malibu have much higher price tags.

What Do You Think?

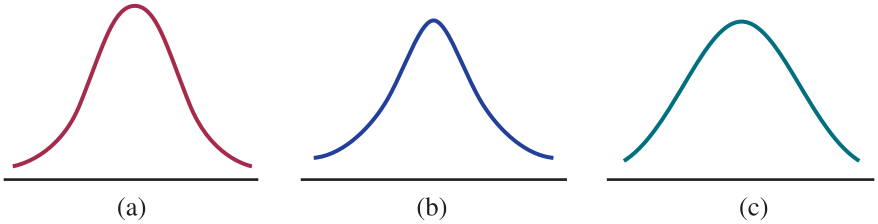
What is another example of a data distribution that you think would be skewed? Would it be positively or negatively skewed?

Figure 3.31
Two examples of skewed smoothed histograms:
(a) positive skew;
(b) negative skew.



One specific shape, a **normal curve**, arises more frequently than any other in statistical applications. Many histograms can be well approximated by a normal curve (for example, characteristics such as arm span and the weight of an apple). Several of the most important characteristics of normal curves are mentioned here. A more detailed discussion follows in Chapter 7. A normal curve is both symmetric and bell-shaped, and it looks like the curve in Figure 3.32(a). However, not all bell-shaped curves are normal. In a normal curve, starting from the top of the bell the height of the curve decreases at a well-defined rate when moving toward either tail. (This rate of decrease is specified by a certain mathematical function.)

Figure 3.32
Three examples of bell-shaped histograms:
(a) normal;
(b) heavy-tailed;
(c) light-tailed.



A curve with tails that do not decline as rapidly as the tails of a normal curve is called **heavy-tailed** (compared to the normal curve). Similarly, a curve with tails that decrease more rapidly than the normal tails is called **light-tailed**. Figure 3.32(b) and (c) illustrate these possibilities. The reason the tails in a distribution are of interest is that many inferential procedures that work well (meaning they result in a high proportion of correct conclusions) when the population distribution is approximately normal perform poorly when the population distribution is heavy-tailed.

Copyright 2025 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s). Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

Do Sample Histograms Resemble Population Histograms?

Sample data are usually collected to learn about a population. Conclusions based on sample data may not be correct if the sample is unrepresentative of the population. So how similar might a histogram of sample data be to the histogram of all population values? Will the two histograms be similar in terms of center and variability? Will they have the same number of peaks, and will the peaks occur at approximately the same places?

A related issue concerns the extent to which histograms based on different samples from the same population resemble one another. If two different sample histograms can be expected to differ from one another in obvious ways, then at least one of them might differ substantially from the population histogram. If the sample differs substantially from the population, conclusions about the population based on the sample are likely to be incorrect. **Sampling variability**—the extent to which samples from the same population differ from one another and from the population—is a central idea in statistics. Example 3.19 considers sampling variability in histogram shapes.

Example 3.19 What You Should Know About Bus Drivers . . .

Understand the context ➤

● A sample of 708 bus drivers employed by public corporations was selected, and the number of traffic accidents in which each bus driver was involved during a 4-year period was determined **("Application of Discrete Distribution Theory to the Study of Noncommunicable Events in Medical Epidemiology,"** ***Random Counts in Biomedical and Social Sciences,*** **G. P. Patil, ed., 1970)**. A listing of the 708 sample observations might look like this:

3 0 6 0 0 2 1 4 1 ... 6 0 2

The frequency distribution (Table 3.7) shows that 117 of the 708 drivers had no accidents, a relative frequency of 117/708 = 0.165 (or 16.5%). Similarly, the proportion of sampled drivers who had 1 accident is 0.222 (or 22.2%). The largest sample observation was 11.

Table 3.7 Frequency Distribution for Number of Accidents by Bus Drivers

Number of Accidents	Frequency	Relative Frequency
0	117	0.165
1	157	0.222
2	158	0.223
3	115	0.162
4	78	0.110
5	44	0.062
6	21	0.030
7	7	0.010
8	6	0.008
9	1	0.001
10	3	0.004
11	1	0.001
	708	0.998

Although the 708 observations were actually a sample from the population of all bus drivers, in this example the 708 observations will be considered as a population. The first histogram in Figure 3.33, represents the population histogram based on all 708 observations. The other four histograms in Figure 3.33 are based on four different random samples of 50 observations from this population. The five histograms resemble one another in a general way, but there are also some obvious differences. The population histogram rises to a peak and then declines smoothly, whereas the sample histograms tend to have more peaks, valleys, and gaps. Although the population data set contained an observation of 11, none of the four samples did. In fact, in the first two samples, the largest observations were 7 and 8, respectively. In Chapters 8–15 you will see how sampling variability can be described and taken into account when sample data is used to draw conclusions about a population.

● **Data set available online**

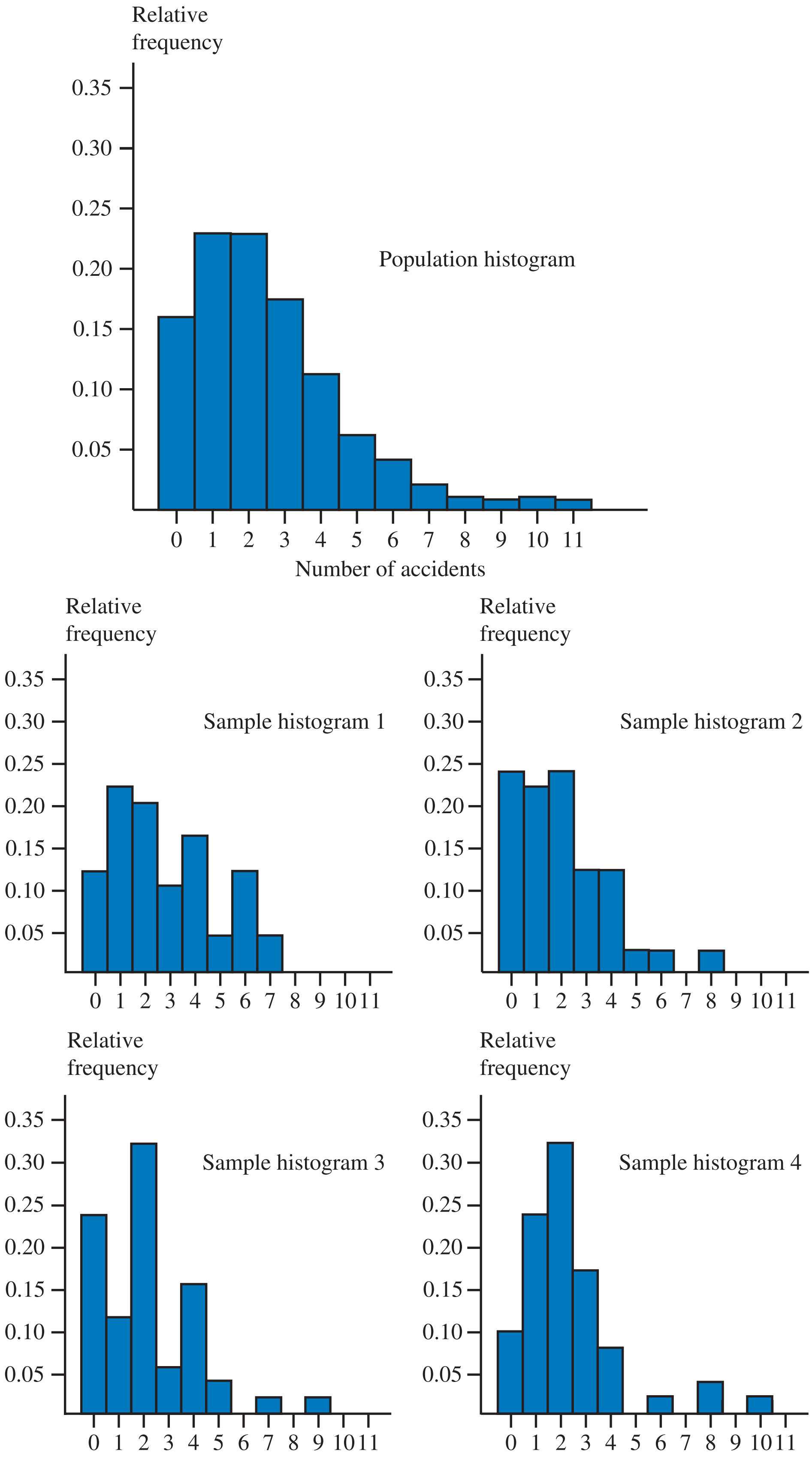


Figure 3.33
Comparison of population and sample histograms for number of accidents.

Cumulative Relative Frequencies and Cumulative Relative Frequency Plots

Rather than wanting to know what proportion of the data fall in a particular class interval, you may want to determine the proportion falling below a specified value. This is easily done when the value is a class interval boundary.

Consider the following intervals and relative frequencies for carry-on luggage weight (in lbs.) for passengers on flights between Phoenix and New York City:

Weight	0 to 5	5 to < 10	10 to < 15	15 to < 20	...
Relative frequency	0.05	0.10	0.18	0.25	...

Then

proportion of passengers with carry-on luggage weight less than 15 lbs.

$$\begin{aligned} &= \text{proportion in one of the first three intervals} \\ &= 0.05 + 0.10 + 0.18 \\ &= 0.33 \end{aligned}$$

Similarly,

proportion of passengers with carry-on luggage weight less than 20 lbs.

$$= 0.05 + 0.10 + 0.18 + 0.25 = 0.33 + 0.25 = 0.58$$

Each such sum of relative frequencies is called a **cumulative relative frequency**. Notice that the cumulative relative frequency 0.58 is the sum of the previous cumulative relative frequency 0.33 and the "current" relative frequency 0.25. The use of cumulative relative frequencies is illustrated in Example 3.20.

Example 3.20 Albuquerque Rainfall

Understand the context ➤

● **The National Climatic Data Center** has been collecting weather data for many years. Annual rainfall totals for Albuquerque, New Mexico, from 1992 to 2021 **(ncdc.noaa.gov)** were used to construct the relative frequency distribution shown in Table 3.8. The table also contains a column of cumulative relative frequencies.

Table 3.8 Relative Frequency Distribution for Albuquerque Rainfall Data with Cumulative Relative Frequencies

Annual Rainfall (inches)	Frequency	Relative Frequency	Cumulative Relative Frequency
4 to < 5	1	0.033	0.033
5 to < 6	4	0.133	0.166 = 0.033 + 0.133
6 to < 7	5	0.167	0.333 = 0.033 + 0.133 + 0.167 or 0.133 + 0.167
7 to < 8	1	0.033	0.366
8 to < 9	7	0.233	0.599
9 to < 10	4	0.133	0.732
10 to < 11	1	0.033	0.765
11 to < 12	4	0.133	0.898
12 to < 13	2	0.067	0.965
13 to < 14	1	0.033	0.998

The proportion of years with annual rainfall less than 10 inches is 0.732, the cumulative relative frequency for the 9 to < 10 interval. What about the proportion of years with annual rainfall less than 8.5 inches? Because 8.5 is not the endpoint of one of the intervals in the frequency distribution, you can only estimate this from the information given. The value 8.5 is halfway between the endpoints of the 8 to 9 interval, so it is reasonable

● **Data set available online**

Copyright 2025 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s). Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

to estimate that half of the relative frequency of 0.233 for this interval belongs in the 8 to 8.5 range. Then

$$\begin{pmatrix}\text{estimate of proportion of}\\\text{years with rainfall less}\\\text{than 8.5 inches}\end{pmatrix} = 0.033 + 0.133 + 0.167 + 0.033 + \frac{1}{2}(0.233) = 0.483$$

This proportion could also have been estimated using the cumulative relative frequencies as

$$\begin{pmatrix}\text{estimate of proportion of}\\\text{years with rainfall less}\\\text{than 8.5 inches}\end{pmatrix} = 0.366 + \frac{1}{2}(0.233) = 0.483$$

Similarly, since 11.25 is one-fourth of the way between 11 and 12,

$$\begin{pmatrix}\text{estimate of proportion of}\\\text{years with rainfall less}\\\text{than 11.25 inches}\end{pmatrix} = 0.765 + \frac{1}{4}(0.133) = 0.798$$

A **cumulative relative frequency plot** is a graph of the cumulative relative frequencies against the upper endpoint of the corresponding interval. The pairs

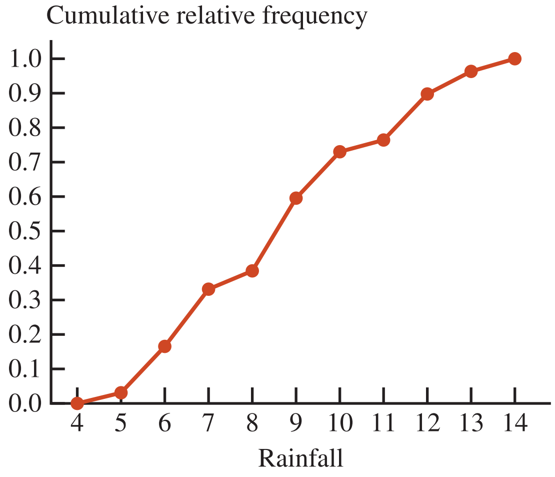
(upper endpoint of interval, cumulative relative frequency)

are plotted as points in two dimensions with the upper endpoint of the interval on the x axis and the cumulative relative frequency on the y axis. Then, successive points in the plot are connected by a line segment. For the rainfall data of Example 3.19, the plotted points would be

(5, 0.033)	(6, 0.166)	(7, 0.333)	(8, 0.366)	(9, 0.599)
(10, 0.732)	(11, 0.765)	(12, 0.898)	(13, 0.965)	(14, 0.998)

One additional point, the pair (lower endpoint of first interval, 0), is also included in the plot (for the rainfall data, this would be the point (4, 0)). Then, points are connected by line segments. Figure 3.34 shows the cumulative relative frequency plot for the rainfall data. The cumulative relative frequency plot can be used to obtain approximate answers to questions such as

Figure 3.34
Cumulative relative frequency plot for the rainfall data of Example 3.20.



What proportion of the observations is smaller than a particular value?

and

What value separates the smallest *p* percent from the larger values?

For example, to determine the approximate proportion of years with annual rainfall less than 9.5 inches, you would follow a vertical line up from 9.5 on the *x*-axis and then read across to the *y*-axis to obtain the corresponding relative frequency, as illustrated in Figure 3.35(a). Approximately 0.68, or 68%, of the years had annual rainfall less than 9.5 inches.

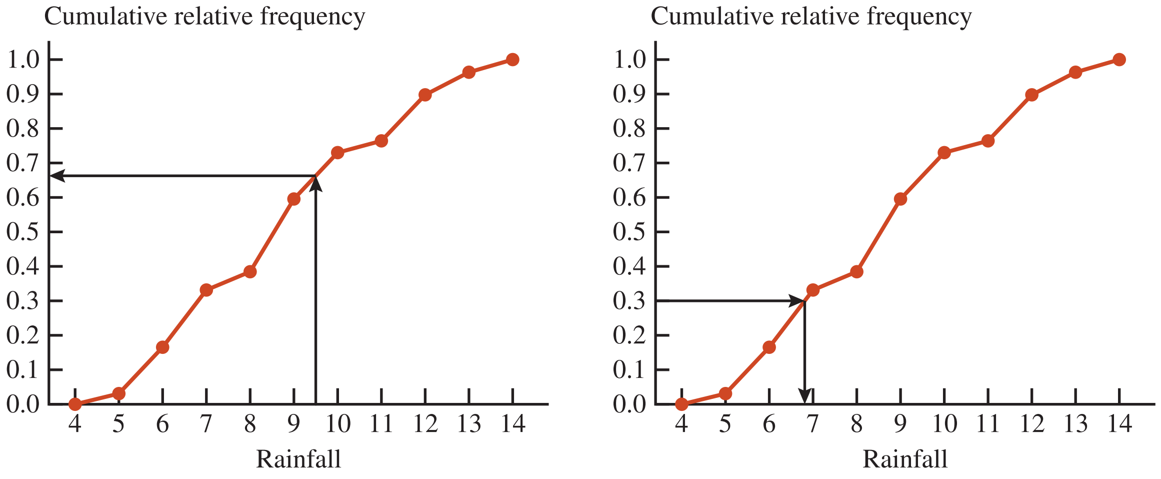
Similarly, to find the approximate amount of rainfall that separates the 30% of years with the smallest annual rainfall from years with higher rainfall, start at 0.30 on the cumulative relative frequency axis and move across and then down to find the corresponding rainfall amount, as shown in Figure 3.35(b). Approximately 30% of the years had annual rainfall of less than about 6.75 inches.

Figure 3.35

Using the cumulative relative frequency plot.

(a) Determining the approximate proportion of years with annual rainfall less than 9.5 inches.

(b) Finding the amount of rainfall that separates the 30% of years with the lowest rainfall from the 70% with higher rainfall.



Exercises 3.24–3.39

● **Data set available online**

3.24 ● The data in the accompanying table are from the **Organisation for Economic Co-operation and Development (data.oecd.org).** The entries in this table are the percentage of 25- to 34-year-old people who have completed a 4-year college degree for 43 countries in 2019.

Country	Percentage of people age 25 to 34 with 4-year degree	Country	Percentage of people age 25 to 34 with 4-year degree
Argentina	40.0	Costa Rica	31.4
Australia	52.5	Czech Republic	32.6
Austria	41.6	Denmark	47.1
Belgium	47.3	Estonia	42.8
Brazil	21.3	Finland	41.8
Canada	63.0	France	48.1
Chile	33.7	Germany	33.3
Columbia	30.0		

(continued)

Country	Percentage of people age 25 to 34 with 4-year degree	Country	Percentage of people age 25 to 34 with 4-year degree
Greece	42.4	Norway	48.7
Hungary	30.6	Poland	43.5
Iceland	47.1	Portugal	37.4
Indonesia	16.1	Russia	62.1
Ireland	55.5	Slovak Republic	39.2
Israel	47.0	Slovenia	44.1
Italy	25.6	South Africa	5.6
Japan	60.1	Spain	46.5
Korea	69.8	Sweden	48.4
Latvia	43.8	Switzerland	52.7
Lithuania	55.2	Turkey	35.3
Luxembourg	56.0	United Kingdom	51.8
Mexico	21.8	United States	50.4
Netherlands	49.1		
New Zealand	43.8		

Copyright 2025 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s). Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

a. Construct a histogram of these data using the class intervals 0 to < 10, 10 to < 20, … , 60 to < 70.

b. Write a few sentences describing the shape, center, and variability of the distribution.

c. Are there any unusual values in the data set? What do you notice about these values?

3.25 The authors of the article **"Dirty Money: New Research Reveals the Filthiness of Our Cash, Cards, & Coins" (lendedu.com, 2020)** measured the bacteria on the surface of credit and debit cards, bills, and coins. Germ scores (with higher scores corresponding to dirtier surfaces) for 41 credit and debit cards are summarized in the following frequency distribution.

Germ Score	Frequency
0 to < 100	4
100 to < 200	23
200 to < 300	3
300 to < 500	5
500 to < 1000	4
1000 to < 1500	2

Summarize this data set using a histogram. Because the class intervals are not all the same width, you will need to use a density scale for the histogram.

3.26 Each year, the College Board reports scores from its Advanced Placement (AP) exams (**apstudents.collegeboard.org**). A student taking an AP exam receives a score of 1, 2, 3, 4, or 5. The accompanying table shows the relative frequency score distributions for the 2020 AP Computer Science Principles exam and the AP Statistics exam.

Score	Relative Frequency Computer Science Principles	Relative Frequency Statistics
5	0.109	0.162
4	0.236	0.207
3	0.371	0.231
2	0.198	0.217
1	0.086	0.183

a. Construct a relative frequency histogram for the AP Computer Science Principles scores.

b. Using the same scales as the histogram from Part (a), construct a relative frequency histogram for the AP Statistics scores.

c. Based on the histograms from Parts (a) and (b), write a few sentences commenting on the similarities and differences in the two score distributions.

3.27 ● The data in the accompanying table represents the percentage of workers in 2022 who were members of a union for each U.S. state and the District of Columbia **(Bureau of Labor Statistics, January 2023)**

State	% of Workers who Belong to a Union
Alabama	8.4
Alaska	17.3
Arizona	6.2
Arkansas	5.7
California	17.6
Colorado	7.5
Connecticut	15.4
Delaware	9.4
District of Columbia	10.4
Florida	5.5
Georgia	3.7
Hawaii	23.4
Idaho	5.9
Illinois	14.1
Indiana	8.6
Iowa	8.7
Kansas	12.1
Kentucky	10.3
Louisiana	5.2
Maine	11.5
Maryland	13.2
Massachusetts	13.8
Michigan	15.3
Minnesota	15.2
Mississippi	7.2
Missouri	10.6
Montana	12.3
Nebraska	8.1
Nevada	12.8
New Hampshire	11.2
New Jersey	16.0
New Mexico	10.6
New York	22.1
North Carolina	3.1
North Dakota	7.6
Ohio	14.0
Oklahoma	7.1
Oregon	16.9
Pennsylvania	13.6
Rhode Island	17.7
South Carolina	2.0
South Dakota	4.2
Tennessee	6.3
Texas	5.1
Utah	8.7
Vermont	13.4
Virginia	4.5
Washington	19.1
West Virginia	10.0
Wisconsin	8.0
Wyoming	7.3

a. Construct a histogram of these data using class intervals of 0 to < 5, 5 to < 10, 10 to < 15, 15 to < 20, and 20 to < 25.

b. Construct a dotplot of these data. Comment on the interesting features of the plot. (Hint: Dotplots were covered in Section 1.4.)

c. For this data set, which is a more informative graphical display—the dotplot from Part (b) or the histogram constructed in Part (a)? Explain.

3.28 Construct a histogram for the data in the previous exercise using about twice as many class intervals. Use 0 to < 2.5 as the first class interval. Write a few sentences that explain why this histogram does a better job of displaying this data set than the histogram in the previous exercise.

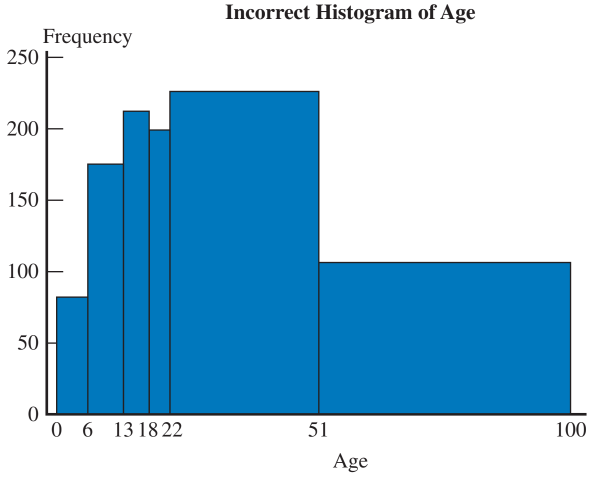
3.29 The paper **"Simpson's Paradox: A Data and Discrimination Case Study Exercise" (*Journal of Statistics Education* [2014])** included data on the ages (in years) of clients of the California Department of Developmental Services. The authors grouped the age data using intervals based on typical costs for

Copyright 2025 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s). Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

services, using the age groups in the frequency distribution below.

Age Interval	Frequency
0 to < 6	82
6 to < 13	175
13 to < 18	212
18 to < 22	199
22 to < 51	226
51 to < 100	106

Below is an **incorrectly** drawn histogram that uses frequency on the vertical scale.



Construct a correctly drawn histogram that uses density to determine the heights of the rectangles in the histogram. Compare the correct histogram to the one given here and comment on how the incorrectly drawn histogram is misleading.

3.30 ● U.S. Census data for San Luis Obispo County, California, were used to construct the following relative frequency distribution for commute time (in minutes) of working adults in 2015.

Commute Time	Relative Frequency
0 to < 5	0.056
5 to < 10	0.156
10 to < 15	0.177
15 to < 20	0.155
20 to < 25	0.147
25 to < 30	0.061
30 to < 35	0.121
35 to < 40	0.015
40 to < 45	0.024
45 to < 60	0.040
60 to < 90	0.030
90 to < 120	0.018

a. Notice that not all intervals in the frequency distribution are equal in width. Why do you think that unequal width intervals were used?

b. Construct a table that adds a density column to the given relative frequency distribution. (Hint: See Example 3.18.)

c. Use the densities computed in Part (b) to construct a histogram for this data set. (The website that was the source of these data actually displays an incorrectly drawn histogram based on relative frequencies rather than densities!) Write a few sentences commenting on the important features of the histogram.

3.31 Use the commute time data given in the previous exercise to complete the following:

a. Calculate the cumulative relative frequencies, and construct a cumulative relative frequency plot.

b. Use the cumulative relative frequency plot constructed in Part (a) to answer the following questions.

i. Approximately what proportion of commute times were less than 50 minutes?

ii. Approximately what proportion of commute times were greater than 22 minutes?

iii. What is the approximate commute time value that separates the shortest 50% of commute times from the longest 50%?

3.32 The report **"Trends in College Pricing and Student Aid 2022" (collegeboard.com)** included data on the average cost of in-state tuition and fees for public two- and four-year colleges for each of the 50 U.S. states. These data were used to construct the accompanying frequency distributions.

Tuition and Fees	Four-Year Public Colleges	Two-Year Public Colleges
0 to < 2,500	0	3
2,500 to < 5,000	0	30
5,000 to < 7,500	4	16
7,500 to < 10,000	15	1
10,000 to < 12,500	16	0
12,500 to < 15,000	10	0
15,000 to < 17,500	5	0

a. Construct a histogram for tuition and fees for four-year public colleges. Write a few sentences describing the distribution of tuition and fees, commenting on center, variability, and shape.

b. Construct a histogram for tuition and fees for two-year public colleges. Write a few sentences describing the distribution of tuition and fees, commenting on center, variability, and shape.

c. Write a few sentences describing the differences in the distributions.

Copyright 2025 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s). Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

3.33 An exam is given to students in an introductory statistics course. What is likely to be true of the shape of the histogram of scores if:

a. the exam is quite easy?
b. the exam is quite difficult?
c. half the students in the class have had calculus, the other half have had no prior college math courses, and the exam emphasizes mathematical manipulation?

Explain your reasoning in each case.

3.34 The following two relative frequency distributions were constructed from data in the **"Majoring in Money"** reports from 2016 and 2019 **(salliemae.com)**. One distribution summarizes data for a sample of 445 college students in 2016 and the other summarizes data from a sample of 464 college students in 2019.

Credit Card Balance (dollars)—2019 Sample	Relative Frequency
0 to < 100	0.26
100 to < 500	0.36
500 to < 1000	0.14
1000 to < 2500	0.13
2500 to < 5000	0.08
5000 or more	0.03

Credit Card Balance (dollars)—2016Sample	Relative Frequency
0 to < 100	0.30
100 to < 500	0.39
500 to < 1000	0.15
1000 to < 2500	0.05
2500 to < 5000	0.08
5000 or more	0.03

a. Construct a histogram for the 2019 survey data. Assume that no one had a balance greater than 15,000 and that the last interval is 5000 to < 15,000. Be sure to use the density scale.
b. Construct a histogram for the 2016 survey data. Use the same scales that you used for the histogram in Part (a) so that it will be easy to compare the two.
c. Comment on the similarities and differences in the histograms from Parts (a) and (b).
d. Based on the two histograms, do you think that student credit card balances tended to be higher in 2019 than they were in 2016? Explain.

3.35 ● Example 3.20 used annual rainfall data for Albuquerque, New Mexico, to construct a relative frequency distribution and cumulative relative frequency plot. The National Climate Data Center also gave the accompanying annual rainfall (in inches) for Medford, Oregon, from 1992 to 2021.

14.98	16.60	12.10	21.77	31.41	17.93	28.72	16.50	18.81	15.01
18.00	19.84	19.09	23.42	21.80	17.10	13.81	11.81	21.31	16.42
26.87	8.99	20.25	17.48	21.71	18.13	12.06	20.33	15.43	16.14

a. Construct a relative frequency distribution for the Medford rainfall data.
b. Use the relative frequency distribution of Part (a) to construct a histogram. Describe the shape of the histogram.

3.36 Use the relative frequency distribution constructed in the previous exercise to answer the following questions.

a. Construct a cumulative relative frequency plot for the Medford rainfall data.
b. Use the cumulative relative frequency plot of Part (a) to answer the following questions:
  i. Approximately what proportion of years had annual rainfall less than 15.5 inches?
  ii. Approximately what proportion of years had annual rainfall less than 25 inches?
  iii. Approximately what proportion of years had annual rainfall between 17.5 and 25 inches?

3.37 The authors of the paper **"Myeloma in Patients Younger than Age 50 Years Presents with More Favorable Features and Shows Better Survival"** (***Blood*** **[2008]: 4039–4047)** studied patients who had been diagnosed with stage 2 multiple myeloma prior to the age of 50. For each patient who received high dose chemotherapy, the number of years that the patient lived after the therapy (survival time) was recorded. The cumulative relative frequencies in the accompanying table were approximated from survival graphs that appeared in the paper.

Years Survived	Cumulative Relative Frequency
0 to < 2	0.10
2 to < 4	0.52
4 to < 6	0.54
6 to < 8	0.64
8 to < 10	0.68
10 to < 12	0.70
12 to < 14	0.72
14 to < 16	1.00

a. Use the given information to construct a cumulative relative frequency plot.
b. Use the cumulative relative frequency plot from Part (a) to answer the following questions:
  i. What is the approximate proportion of patients who lived fewer than 5 years after treatment?

ii. What is the approximate proportion of patients who lived fewer than 7.5 years after treatment?

iii. What is the approximate proportion of patients who lived more than 10 years after treatment?

d. What additional information would you need in order to decide if it is reasonable to generalize conclusions about survival time from the group of patients in the study to all patients younger than 50 years old who are diagnosed with multiple myeloma and who receive high dose chemotherapy?

3.38 Use the cumulative relative frequencies given in the previous exercise to complete the following:

a. Calculate the relative frequencies for each class interval and construct a relative frequency distribution.

b. Summarize the survival time data using a histogram.

c. Based on the histogram, write a few sentences describing survival time of the stage 2 myeloma patients in this study.

3.39 Using the five class intervals 100 to 120, 120 to 140, . . . , 180 to 200, devise a frequency distribution based on 70 observations whose histogram could be described as follows:

a. symmetric

b. bimodal

c. positively skewed

d. negatively skewed

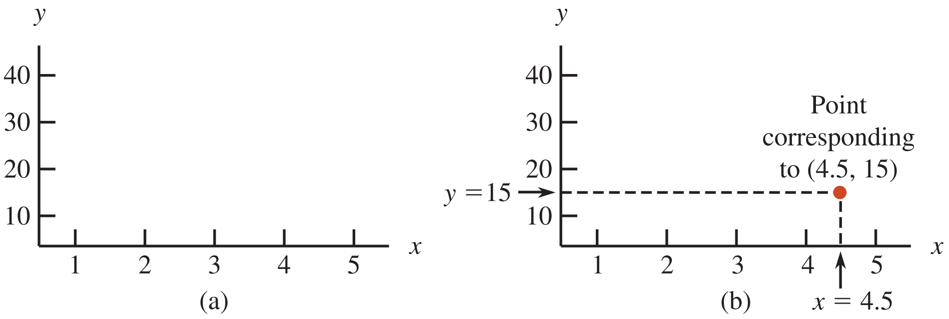
Section 3.4 Displaying Bivariate Numerical Data

A bivariate data set consists of measurements or observations on two variables, x and y. For example, x might be the distance from a highway and y the lead content of soil at that distance. When both x and y are numerical variables, each observation consists of a pair of numbers, such as (14, 5.2) or (27.63, 18.9). The first number in a pair is the value of x, and the second number is the value of y.

An unorganized list of bivariate data provides little information about the distribution of either the x values or the y values separately and even less information about how the two variables are related to one another. Just as graphical displays can be used to summarize univariate data, they can also be used to summarize bivariate data. The most important graph based on bivariate numerical data is a **scatterplot**.

In a scatterplot each observation (pair of numbers) is represented by a point on a rectangular coordinate system, as shown in Figure 3.36(a). The horizontal axis is identified with values of x and is scaled so that any x value can be easily located. Similarly, the vertical or y-axis is marked for easy location of y values. The point corresponding to any particular (x, y) pair is placed where a vertical line from the value on the x-axis meets a horizontal line from the value on the y-axis. Figure 3.36(b) shows the point representing the observation (4.5, 15). It is above 4.5 on the horizontal axis and to the right of 15 on the vertical axis.

Figure 3.36
Constructing a scatterplot:
(a) rectangular coordinate system;
(b) point corresponding to (4.5, 15).



Example 3.21 Olympic Figure Skating

Understand the context ➤

● Do tall skaters have an advantage when it comes to earning high artistic scores in figure skating competitions? Data on x = height (in inches) and y = components score (based on artistry, interpretation, and presentation) in the free skate for both men's and women's singles skaters at the 2018 Winter Olympics are shown in the accompanying table. (Data from **olympic.org**)

● **Data set available online**

Copyright 2025 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s). Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

Men's Singles Skater	**Height (in.)**	**Components Score**	**Women's Singles Skater**	**Height (in.)**	**Components Score**
Yuzuru HANYU	67	96.62	Alina ZAGITOVA	61	75.03
Shoma UNO	63	92.72	Evgenia MEDVEDEVA	63	77.47
Javier FERNANDEZ	68	96.14	Kaetlyn OSMOND	64	75.65
JIN Boyang	67	85.76	Satoko MIYAHARA	60	71.24
Nathan CHEN	65	87.44	Carolina KOSTNER	67	75.65
Vincent ZHOU	67	79.92	Kaori SAKAMOTO	62	68.11
Dmitri ALIEV	70	85.14	CHOI Dabin	61	62.75
Mikhail KOLYADA	69	87.94	Maria SOTSKOVA	68	67.30
Patrick CHAN	66	91.86	Bradie TENNELL	66	62.93
Adam RIPPON	67	86.94	Mirai NAGASU	63	62.05
Alexei BYCHENKO	69	83.80	Karen CHEN	60	64.10
Keegan MESSING	64	85.44	Elizabet TURSYNBAEVA	59	58.80
Daniel SAMOHIN	69	81.72	KIM Hanul	66	54.35
Jorik HENDRICKX	69	82.42	Nicole RAJICOVA	64	56.80
CHA Junhwan	69	81.22	Gabrielle DALEMAN	60	61.75
Michal BREZINA	68	84.34	Loena HENDRICKX	63	55.99
Misha GE	68	86.08	Kailani CRAINE	63	53.95
Keiji TANAKA	68	81.14	Nicole SCHOTT	65	56.58
Deniss VASILJEVS	68	80.64	Mae Berenice MEITE	66	52.12
Brendan KERRY	68	77.42	Emmi PELTONEN	64	56.45
Matteo RIZZO	67	75.92	Alexia PAGANINI	66	50.06
Paul FENTZ	70	72.84	LI Xiangning	57	51.41
YAN Han	67	79.58	Ivett TOTH	61	50.39
Morisi KVITELASHVILI	71	70.66	Isadora WILLIAMS	61	51.05

Consider the data ➤

Figure 3.37(a) gives a scatterplot of the data. Looking at the data and the scatterplot, you can see that

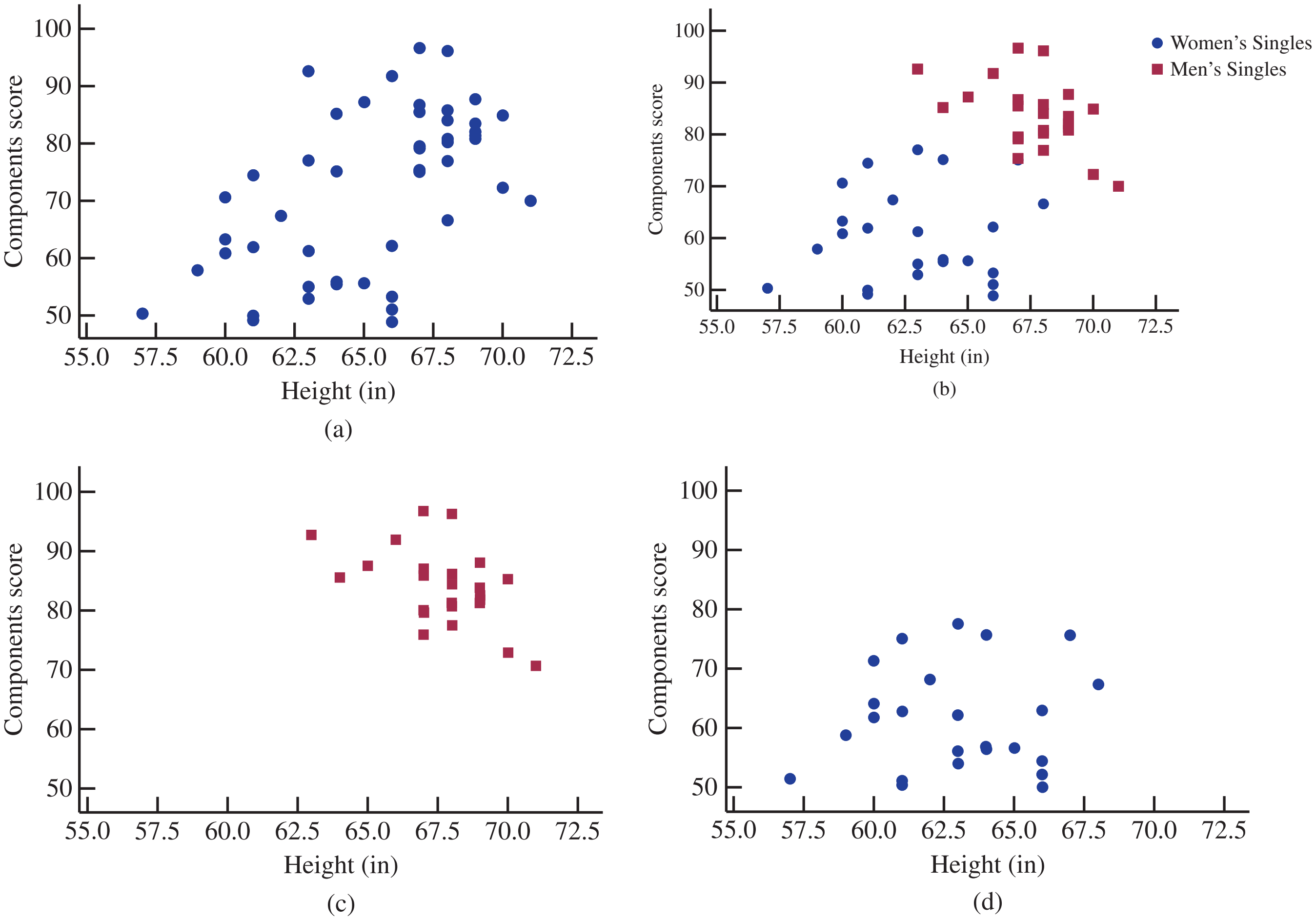
Interpret the results ➤

1. Several observations have identical x values but different y values (for example, $x = 67$ inches for both Yuzuru Hanyu and Adam Rippon, but Hanyu's component score was 96.62 and Rippon's component score was 86.94). This shows that the value of y is *not* determined *solely* by the value of x but by various other factors as well.
2. At any given height there is quite a bit of variability in artistic score. For example, for those skaters with height 67 inches, components scores ranged from a low of about 76 to a high of about 97.
3. There appears to be a tendency for components score to increase as height increases. However, there is not a strong linear relationship between height and components score. The points in the scatterplot do not cluster tightly around a line.

The data set used to construct the scatterplot included data for both men's and women's singles skaters. Figure 3.37(b) shows a scatterplot of the (height, components score) pairs with observations for men's singles skaters shown in red and observations for women's singles skaters shown in blue. Not surprisingly, the women's singles skaters tend to be shorter than the men's singles skaters (the observations for women tend to be concentrated toward the left side of the scatterplot). Careful examination of this plot shows that while there was a weak linear pattern in the combined (men's and women's singles skaters) data set, there does not appear to be a relationship between height and components score for women's singles skaters.

Figures 3.37(c) and (d) show separate scatterplots for the men's and women's singles skaters, respectively. It is interesting to note that it appears that for men's singles skaters, higher components scores seem to be associated with smaller height values, but for women's singles there does not appear to be a relationship between height and components score. The relationship between height and components score for men's singles skaters is not evident in the scatterplot of the combined data, which seems to indicate that higher components scores tend to be paired with greater heights.

Figure 3.37
Scatterplots for the data of Example 3.21: (a) scatterplot of data; (b) scatterplot of data with observations for men and women distinguished by color; (c) scatterplot for men's singles skaters; (d) scatterplot for women's singles skaters.



(a)

(b)

(c)

(d)

The horizontal and vertical axes in the scatterplots of Figure 3.37 do not intersect at the point (0, 0). In many data sets, the values of *x* or of *y* or of both variables differ considerably from 0 relative to the ranges of the values in the data set. For example, a study of how air conditioner efficiency is related to maximum daily outdoor temperature might involve observations at temperatures of 80°, 82°, . . . , 98°, 100°. In such cases, the plot will be more informative if the axes intersect at some point other than (0, 0) and are marked accordingly. This is illustrated in Example 3.22.

Example 3.22 Roller Coaster Speed and Height

Understand the context ➤

● **Data set available online**

● The accompanying table provides data on the top speed (in miles per hour) and the maximum height (in feet) for a sample of 10 roller coasters (**Roller Coaster Data Base, rcdb.com**).

Consider the data ➤

Top Speed (mph)	Maximum Height (feet)
65	150
55	70
50	161
65	150
60	79
65	157
50	84
52	100
80	209
75	200

Figures 3.38(a) and 3.38(b) show two scatterplots of x = maximum height and y = top speed. The scatterplots were produced by the statistical computer package Minitab. In Figure 3.38(a), Minitab selected the scale for both axes. Figure 3.38(b) was obtained by specifying that the axes would intersect at the point (0, 0).

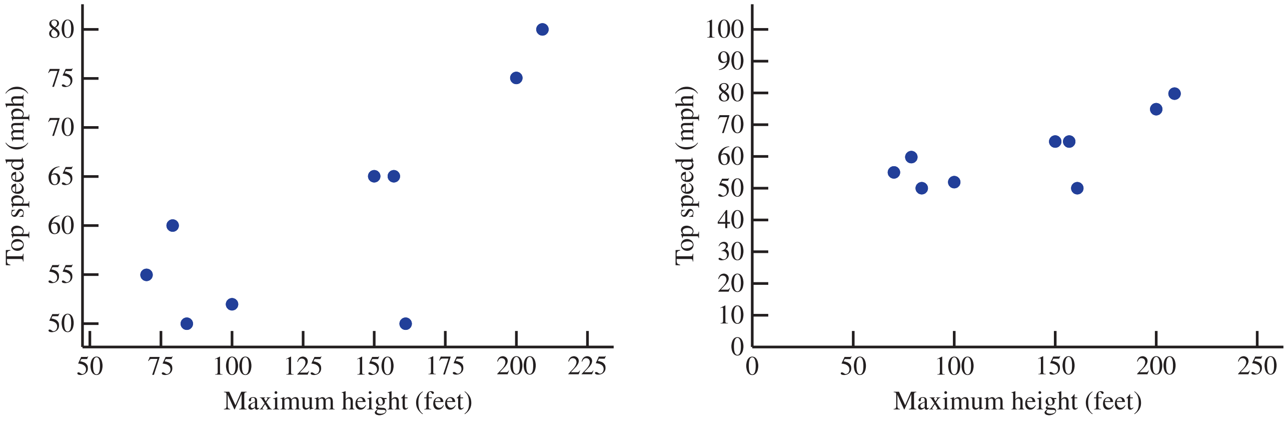
What Do You Think?

Which of the two scatterplots in Figure 3.38 do you think does a better job in communicating the nature of the relationship between top speed and maximum height?

The data points in the second plot are crowded together and it is more difficult to see that top speed tends to increase as maximum height increases. It is also more difficult to see the unusual observation that stands out as being different from the others in the more crowded plot. This observation (the roller coaster with the maximum height of about 160) has a much lower top speed than might have been expected for a roller coaster with this maximum height. This is not as noticeable in the plot where the axes cross at (0,0).

Figure 3.38

Minitab scatterplots of the data in Example 3.22. (a) scale on both axes selected by Minitab; (b) axes intersect at (0,0).



(a)

(b)

Time Series Plots

Data sets often consist of measurements collected over time at regular intervals in order to learn about change over time. For example, stock prices, sales figures, and other socioeconomic indicators might be recorded on a weekly or monthly basis. A **time series plot** (sometimes also called a time plot) is a graph of data collected over time that can be used to identify trends or patterns that might be of interest.

Copyright 2025 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s). Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

A time series plot can be constructed by thinking of the data set as a bivariate data set, where *y* is the variable observed and *x* is the time at which the observation was made. These (*x*, *y*) pairs are plotted as a scatterplot. Consecutive observations are then connected by a line segment. This helps us to see trends over time.

Example 3.23 Wildfires in California

Understand the context ➤

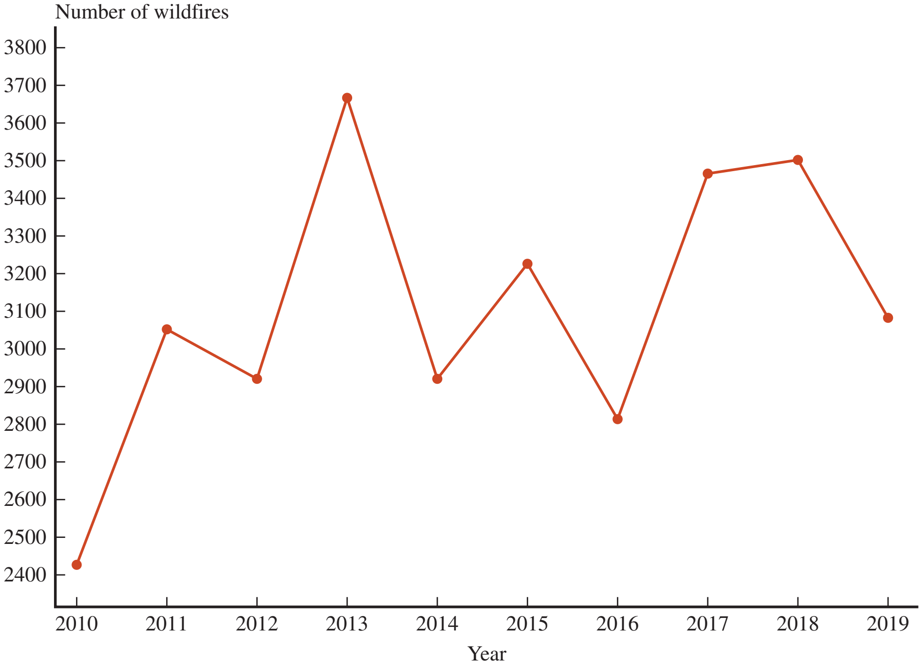
● Each year, Cal Fire reports the number of wildfires occurring in California. Data from the yearly reports for 2010 to 2019 are given in the accompanying table (**www.fire.ca.gov**).

Consider the data ➤

Year	Number of Wildfires
2010	2430
2011	3056
2012	2922
2013	3672
2014	2920
2015	3231
2016	2816
2017	3470
2018	3504
2019	3086

Figure 3.39 shows a time series plot of these data. Notice that the ten (year, number of wildfires) pairs have been plotted and that these points have been connected by line segments. This makes it easier to see any trend over time. You can see from the time series plot that the number of wildfires has not steadily increased year to year, although there does appear to be a general upward trend.

Figure 3.39
Time series plot of the number of wildfires in California.

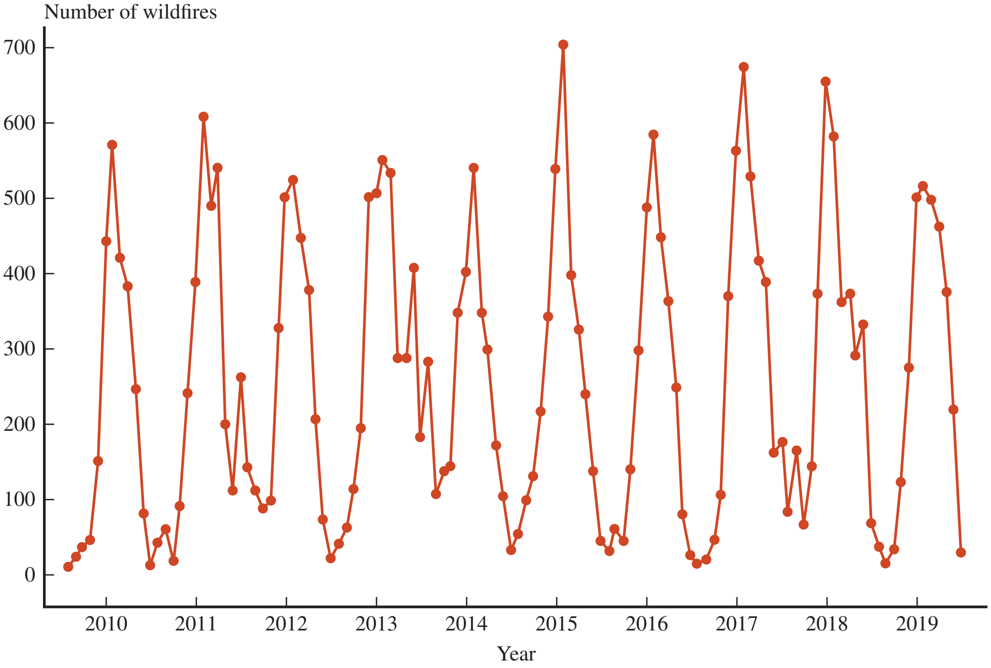


The Cal Fire reports also included data on the number of fires each month over the same 10-year time period. Figure 3.40 is a time series plot based on the monthly data. In addition to the general increasing trend from year to year, you can also see a pattern that repeats each year, with a big increase in the number of fires during the summer months each year.

● **Data set available online**

Copyright 2025 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s). Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

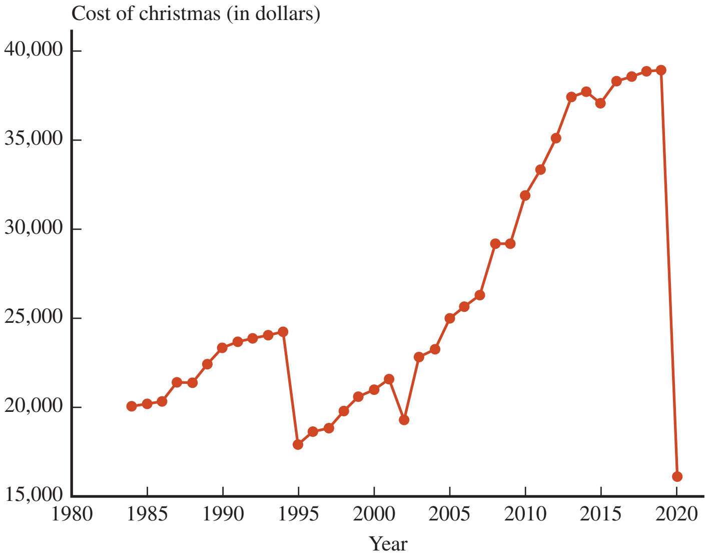
Figure 3.40
Time series plot of number of wildfires in California based on monthly data.



Example 3.24 The Cost of Christmas

Understand the context ➤ The Christmas Price Index is calculated each year by PNC Advisors, and it is a humorous look at the cost of giving all of the gifts described in the popular Christmas song "The Twelve Days of Christmas." The year 2019 was the most costly year since the index began, with the "cost of Christmas" at $38,994. Historical data from the PNC website **(pncchristmaspriceindex.com)** were used to construct the time series plot of Figure 3.41.

Figure 3.41
Time series plot for the Christmas Price Index data of Example 3.24.



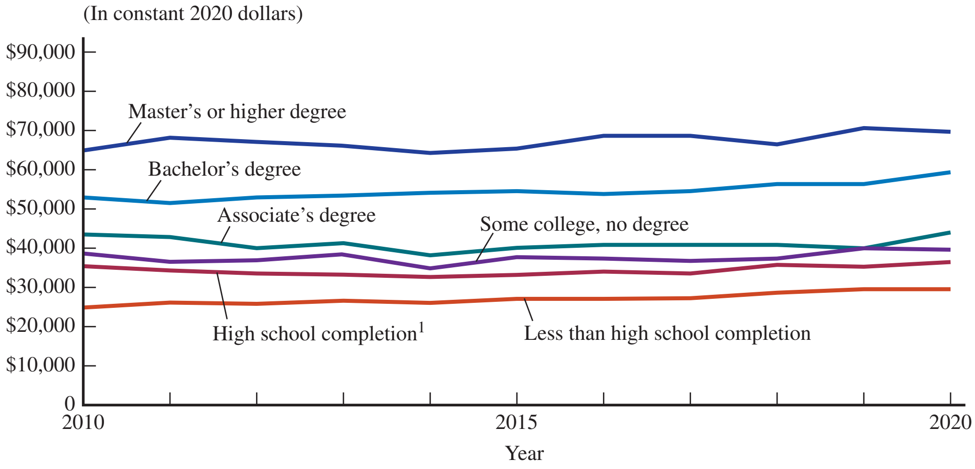
Interpret the results ➤ The plot shows an upward trend in the index from 1984 until 1993. There has been a clear upward trend in the index since 1995. You can visit the website to see individual time series plots for each of the twelve gifts that are used to determine the Christmas Price Index (a partridge in a pear tree, two turtle doves, etc.). The large decrease in 2020 was attributed to the cancellation of live performances in 2020 due to COVID-19–no ladies dancing, no lords-a-leaping, no pipers piping, and no drummers drumming. See if you can figure out what caused the dramatic decline from 1994 to 1995.

Example 3.25 Education and Income—Stay in School!

Consider the data ➤

The time series plot in Figure 3.42 is similar to one appearing on the website of the **National Center for Education Statistics (nces.ed.gov)**. It shows the change over time in median annual earnings by education level. (The median annual earnings is the value for which half of the population earns less and half earns more. For example, in the year 2010, about half of those with bachelor's degrees earned less than $52,000 per year, and half of those with bachelor's degrees earned more than $52,000.) From this plot, you can see that the median annual earnings varied more from year to year for those with master's or higher degrees, but that the large gap between the median earnings for those who only completed high school and those with a college degree has remained about the same since the year 2010.

Figure 3.42
Time series plots of median annual earnings by education level.



Exercises 3.40–3.47

● Data set available online

3.40 ● The accompanying table gives data from a survey of new car owners conducted by the J.D. Power and Associates marketing firm **(The 2020 U.S. Initial Quality Study, jdpowers.com)**. For each brand of car sold in the United States, data on a quality rating (defects per 100 cars, so lower numbers indicate higher quality) and a customer satisfaction rating (called the APEAL rating) are given in the accompanying table. The APEAL rating is a score between 0 and 1000, with higher values indicating greater satisfaction.

Brand	Quality Rating	APEAL Rating
Acura	185	845
Audi	225	845
BMW	176	869
Buick	150	832
Cadillac	162	874
Chevrolet	141	839
Chrysler	189	828
Dodge	136	872
Ford	174	853
Genesis	142	864
GMC	151	857
Honda	177	842
Hyundai	153	842
Infiniti	173	846
Jaguar	190	854
Jeep	155	822
Kia	136	844
Land Rover	228	866
Lexus	159	859
Lincoln	182	876
Mazda	184	844
Mercedes-Benz	202	863
MINI	174	846
Mitsubishi	148	829
Nissan	161	844
Porsche	186	881
Ram	141	871
Subaru	187	832
Tesla	250	896
Toyota	177	825
Volkswagen	152	832
Volvo	210	859

(continued)

Copyright 2025 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s). Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

a. Construct a scatterplot of x = quality rating and y = APEAL rating. (Hint: See Example 3.22).

b. Does customer satisfaction (as measured by the APEAL rating) appear to be related to car quality? Explain.

3.41 ● ***Consumer Reports Health*** **(consumerreports.org)** gave the accompanying data on saturated fat (in grams), sodium (in mg), and calories for 36 fast-food items.

Fat	Sodium	Calories	Fat	Sodium	Calories
2	1042	268	0	520	140
5	921	303	2.5	1120	330
3	250	260	1	240	120
2	770	660	3	650	180
1	635	180	1	1620	340
6	440	290	4	660	380
4.5	490	290	3	840	300
5	1160	360	1.5	1050	490
3.5	970	300	3	1440	380
1	1120	315	9	750	560
2	350	160	1	500	230
3	450	200	1.5	1200	370
6	800	320	2.5	1200	330
3	1190	420	3	1250	330
2	1090	120	0	1040	220
5	570	290	0	760	260
3.5	1215	285	2.5	780	220
2.5	1160	390	3	500	230

a. Construct a scatterplot using y = calories and x = fat. Does it look like there is a relationship between fat and calories? Is the relationship what you expected? Explain.

b. Construct a scatterplot using y = calories and x = sodium. Write a few sentences commenting on the difference between the relationship of calories to fat and calories to sodium.

c. Construct a scatterplot using y = sodium and x = fat. Does there appear to be a relationship between fat and sodium?

d. Add a vertical line at $x = 3$ and a horizontal line at $y = 900$ to the scatterplot in Part (c). This divides the scatterplot into four regions, with some of the points in the scatterplot falling into each of the four regions. Which of the four regions corresponds to healthier fast-food choices? Explain.

3.42 ● The data in the accompanying table are the rate of spread and the wind speed recorded for 32 wildfires that occurred in shrubland vegetation (**"A Generic, Empirical-Based Model for Predicting Rate of Fire Spread in Shrublands,"** ***International Journal of Wildland Fire*** **[2015]: 443–460**).

Rate of Spread (meters per minute)	Wind Speed (km per hour)	Rate of Spread (meters per minute)	Wind Speed (km per hour)
62.5	46.4	16.1	18.0
55.0	54.2	24.3	14.9
8.3	17.9	10.0	12.0
8.0	16.4	50.0	40.0
31.3	26.2	34.0	11.0
6.5	14.9	17.0	11.0
40.2	20.0	75.0	23.0
29.2	28.8	100.0	20.0
12.6	5.0	73.0	20.0
4.8	6.9	19.0	15.0
8.4	18.0	38.0	15.0
36.6	33.0	14.3	18.5
33.6	15.8	19.1	20.0
9.0	11.9	12.0	22.0
17.2	11.9	20.0	23.0
23.3	25.4	15.7	21.0

a. Why would it be of interest to investigate the relationship between wind speed and fire spread?

b. Use the given data to construct a scatterplot of x = wind speed and y = rate of spread.

c. Do any data points stand out as unusual? Explain.

d. The three data points with the highest rates of spread (73, 75, and 100 meters per minute) were all for fires that occurred near the same location in Australia during a one-month period. It was noted that the relative humidity during that month was unusually low. Excluding those three observations from consideration, describe the relationship between wind speed and rate of spread.

3.43 ● ***Consumer Reports*** rated 29 fitness trackers (such as Fitbit and Jawbone) on factors such as ease of use and accuracy of step count to obtain an overall score (**consumerreports.org**). The accompanying table gives price and overall score for these 29 fitness trackers.

Price	Overall Score	Price	Overall Score
260	87	180	74
150	83	125	71
200	82	120	70
200	82	100	69
150	82	90	66
150	79	200	66
130	78	220	65
100	77	60	61
176	74	100	61

(continued)

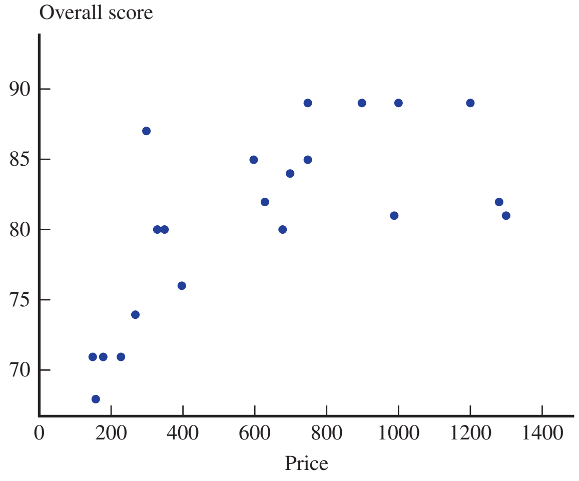
Copyright 2025 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s). Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

Price	**Overall Score**	**Price**	**Overall Score**
100	60	100	54
150	57	125	51
100	56	100	50
70	56	100	44
100	55	22	41
120	54		

a. Construct a scatterplot using y = overall score and x = price.

b. Based on the scatterplot from Part (a), does there appear to be a relationship between price and overall score? Does the scatterplot support the statement that the more expensive fitness trackers tended to receive higher overall scores?

3.44 ***Consumer Reports* (September, 2022)** rated 21 different models of tablet computers that were for sale in 2022. An overall score was assigned to each model based on consideration of several factors, including performance, portability, and battery life. Data on price and overall score were used to construct the following scatterplot.



Would you describe the relationship between price and overall score for these tablet computers as positive (overall score tends to increase as price increases) or negative (overall score tends to decrease as price increases)? Explain.

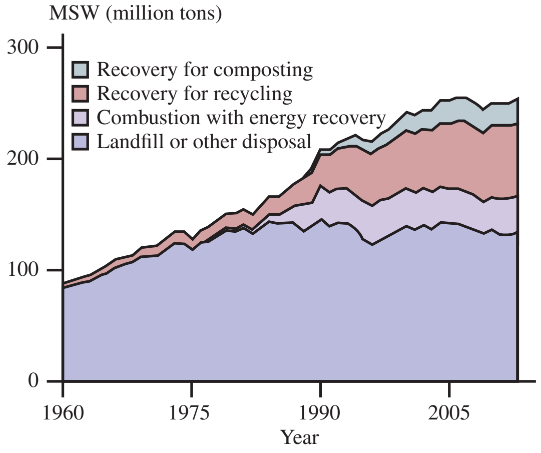
3.45 The Solid Waste Management section of the Environmental Protection Agency ***Report on the Environment* (epa.gov/roe)** included a graph similar to accompanying graph. The report also included the following statement:

> The last several decades have seen steady growth in recycling and composting, while the total amounts landfilled peaked in 1990 (145 MT) and have generally declined since then (134 MT in 2013).

Explain how the time series plot is consistent or is not consistent with the given statement.

Exhibit 1.

Municipal solid waste generated and managed in the U.S., 1960–2013



3.46 ● The report **"The Rate of Children Without Health Insurance is Rising, Particularly among Latino Children of Immigrant Parents and White Children" (National Research Center on Hispanic Children & Families, September 15, 2020)** includes the accompanying data on the percentage of children who were not covered by health insurance for each year from 2010 to 2018.

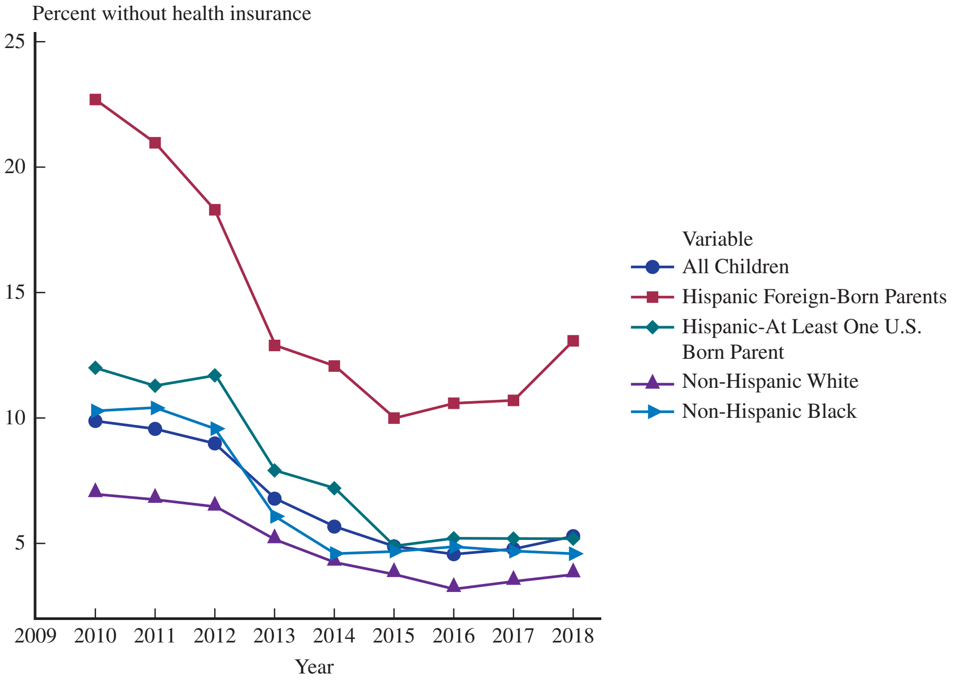
Year	**% Uninsured**
2010	9.9
2011	9.6
2012	9
2013	6.8
2014	5.7
2015	4.9
2016	4.6
2017	4.8
2018	5.3

a. Construct a time series plot and comment on any trend over time.

b. The report also gave data on the percentage of children who were not covered by health insurance for four groups: Hispanic children with both parents foreign-born, Hispanic children with at least one U.S. born parent, Non-Hispanic White children, and Non-Hispanic Black children. Those data were used to create the time series plots for these four groups shown at the top of the next page. The data for all children are also included in the plot. Write a paragraph based on this plot that comments on the similarities and differences of these groups. Also explain whether the title of the report is or is not supported by these time series plots.

Copyright 2025 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s). Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

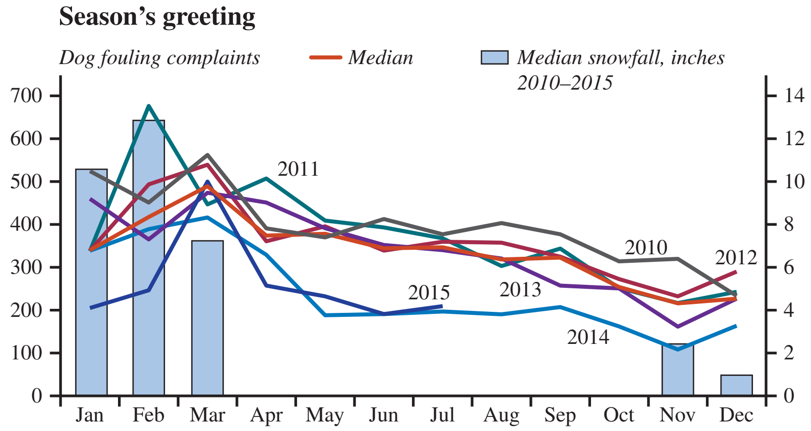
Figure for Exercise 3.46



3.47 The accompanying graph is similar to one appearing in an article titled **"The Seasonality of New York's Dog Poo" (*The Economist*, August 26, 2015)**. The graph includes time series plots of the number of complaints about dog poo on the streets received by the city of New York by month for each of the years from 2010 to 2015. It also includes a bar chart of the median amount of snow in New York City for each month of the year.

a. Patterns that tend to repeat on a regular basis over time are called seasonal patterns. Describe any seasonal patterns that you see in the complaint data. (Hint: Look for patterns that appear to be consistent from year to year.)

b. Why do you think that the authors of this article chose to also include the data on median snowfall in the graph?



Section 3.5 Bivariate and Multivariable Graphical Displays

The use of graphical displays is becoming increasingly common as a way to communicate information. Up to this point, you have considered graphical displays that can be used to summarize univariate and bivariate numerical data. Looking at displays of univariate data, like histograms, dotplots, and bar charts, you can get a sense of what the data distribution looks like. You can also use comparative graphical displays to learn about differences between groups. Looking at a scatterplot of bivariate numerical data allows you to learn about the relationship between two variables. But in your personal and professional life, you will often

Copyright 2025 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s). Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

want to investigate more than two variables to learn about the relationships between them and to understand whether the relationship between two variables might depend on the value of a third variable. For example, you might want to consider if the relationship between life expectancy and income has changed over time or if the relationship differs for different countries. In this section, you will consider a few additional types of graphical displays for bivariate data that allow you to look for relationships between two variables when one of the variables is not numerical. You will also work on developing multivariable thinking by considering several different graphical displays for data sets that contain three or more variables.

Here are a few things to focus on as you consider graphical displays, especially multivariable displays and interactive displays:

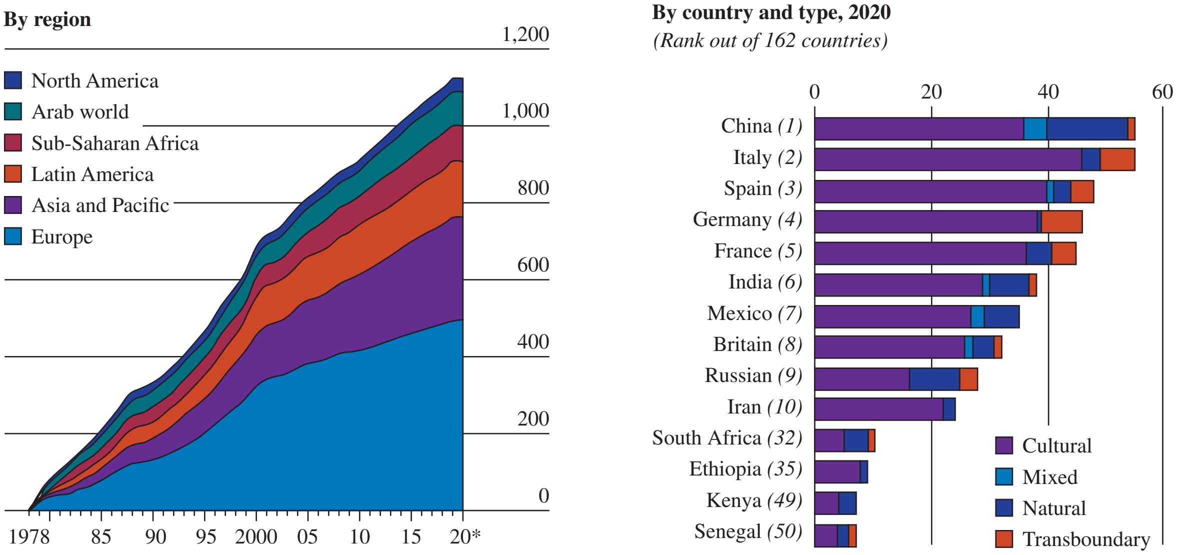
- **What is the context?** Does the display include a title, a legend, or any other text that provides context? Is the source of the data identified? What variables are represented in the graph? Are there labels and axis scales that clarify the context? Is color or shading used in the display, and if so, what do the colors or shading represent?
- **What data distributions and relationships between variables are shown?** What distributions or relationships are shown in the display? Does the distribution of one variable change based on other variables shown in the display? Does the nature of the relationship between two variables depend on the values of other variables that are also included in the display?
- **What story does the graphical display tell?** What have you learned about the context from this display? What does the graph reveal about the relationships between the highlighted variables?
- **Is there room for improvement?** Does the display do a good job communicating its message? Can you think of any ways that the graphical display could be improved?

Keeping these things in mind can help you to think critically when interpreting a graphical display. Now let's consider some examples.

Example 3.26 World Heritage Sites

UNESCO (United Nations Educational, Scientific and Cultural Organization) recognizes sites around the world that are of cultural and historical importance. But very few of these sites are in sub-Saharan Africa. The article **"Why Africa Has Relatively Few UNESCO World Heritage Sites" (*The Economist*, July 22, 2021)** looked at the geographic distribution of the World Heritage sites. The graphical display in Figure 3.43 is similar to the one that appears in the article.

Figure 3.43
UNESCO World Heritage Sites.



Source: UNESCO

*No new sites added in 2020

Copyright 2025 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s). Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

What story does this graphical display tell? There is a lot going on in this display, so one way to start is to consider some of the suggestions given earlier.

- **What is the context?** The title of the graph identifies the context as World Heritage Sites, and the source of the data is identified as UNESCO. There are two parts to this graphical display, and each part has a title. The graph on the left displays data by region and the different regions are represented by different colors which are identified in the given legend. The horizontal axis appears to be year, ranging from 1978 to 2020 (this would have been clearer if the axis had been labeled) and the vertical axis is the total number of World Heritage sites.

  The graph on the right displays data on the distribution of site type and the total number of sites for some of the 162 countries that have World Heritage sites. The different site types are represented by different colors which are identified in the legend. This graph also incorporates information on the rank of each of the countries based on the total number of sites.

  There are several variables represented in this display. To create this display, the data set would need to include the values of the following variables for each World Heritage site: the year the site was added to the list, the region where the site is located, the country where the site is located, the country rank, and the type of site (cultural, mixed, national, or transboundary).
- **What data distributions and relationships between variables are shown?** The graph on the left displays the distribution of sites by region and how this distribution has changed over time. Over time, the number of sites has increased but the graph shows that these increases have not been evenly distributed around the world. The graph on the right compares the number of sites for selected countries. It also shows how the distribution of site type differs from country to country.
- **What story does the graphical display tell?** There are several interesting things that are communicated by this display. One is how much the number of sites in Europe has increased over time compared to the other regions. In the earliest years, there were no big differences in the number of sites in the six regions, but the number of sites has increased at different rates across regions. The number of sites in North America changed the least from 1978 to 2020. The number of sites in the Arab world and sub-Saharan Africa increased over that same period but not as much as the number of sites in Latin America and the Asia Pacific regions increased. However, the increase in the number of sites in Europe was much greater than for any of the other regions.

  Another insight communicated by this graphical display is that for the 10 countries that have the most sites (those with ranks 1–10 in the display on the right), five countries and part of Russia are in Europe and only one is in North America (Mexico). The graph also shows that the top-ranked countries tend to have a large proportion of cultural sites.
- **Is there room for improvement?** This graphical display does do a good job of communicating information about the distribution of World Heritage sites by region over time and also about the differences in the distribution of site type for the ten counties with the most sites and selected sub-Saharan African countries. Two suggestions for improvement would be to add labels to the axes in both of the graphs in this display and to use different colors in the graphical display on the right-hand side of the display than the colors that were used on the left-hand side of the display to avoid possible confusion between regions and site types.

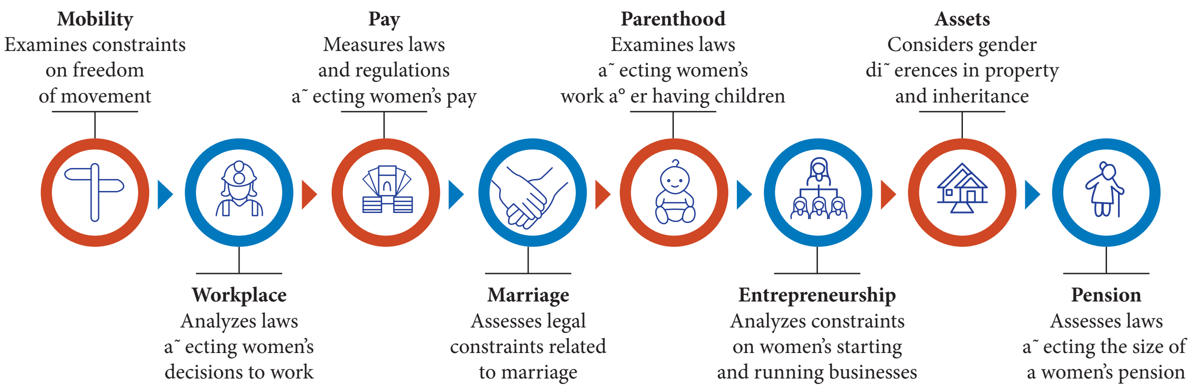
Example 3.27 Women, Business, and the Law

The report “Women, Business and the Law 2021” was prepared by the World Bank **(www.worldbank.org)**. This report looks at factors affecting women’s economic empowerment by considering differences in the law and other policies and practices that support women’s employment. Economies around the world are scored using an index called the

WBL Index (Women, Business and the Law Index). Higher WBL Index values correspond to smaller gender gaps and greater economic opportunity for women. Figure 3.44 is from the report and shows the factors that are considered in constructing the index. (An example of how the index value is calculated for one country (Ecuador) can be found in Table 1.1 of the report.)

Figure 3.44

Components of the Women, Business and the Law Index

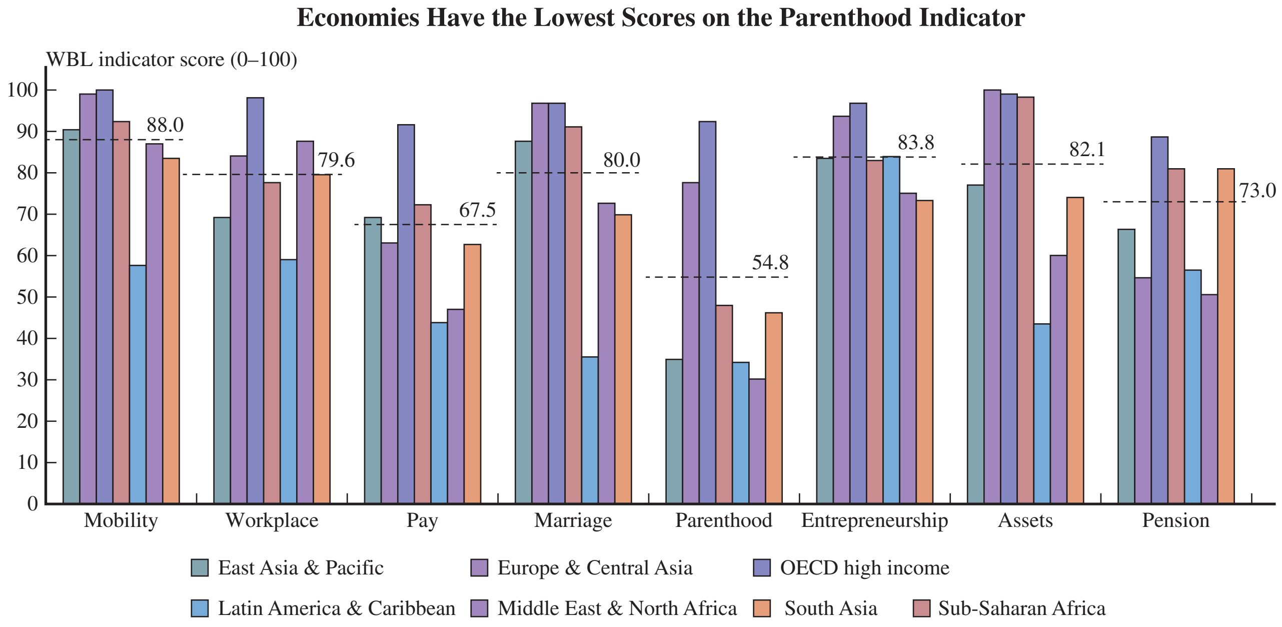

Mobility
Examines constraints on freedom of movement
Workplace
Analyzes laws affecting women's decisions to work
Pay
Measures laws and regulations affecting women's pay
Marriage
Assesses legal constraints related to marriage
Parenthood
Examines laws affecting women's work after having children
Entrepreneurship
Analyzes constraints on women's starting and running businesses
Assets
Considers gender differences in property and inheritance
Pension
Assesses laws affecting the size of a women's pension

Source: *Women, Business and the Law team.*

There are several interesting graphical displays in this report. Two are included in this example and two others are included in the exercises for this section. The graphical display in Figure 3.45 shows the average scores on each component of the WBL Index components for seven different country groups: East Asia & Pacific, Europe and Central Asia, Latin America & Caribbean, Middle East and North Africa, South Africa, Sub-Saharan Africa,

Figure 3.45

WBL Index Component Scores



Source: *Women, Business and the Law* (WBL) database.
Note: OECD = Organisation for Economic Co-operation and Development.

Copyright 2025 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s). Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

and the OECD high-income countries (countries that are members of the Organisation for Economic Co-operation and Development and have a per capita income of greater than $9266, which includes the United States).

What can you learn from the graphical display in Figure 3.45? It is helpful to think about the following:

- **What is the context?** The title of the graphical display is really just a conclusion that can be drawn from the display rather than an overall description of the content of the graph. There are several variables represented in this display: WBL Indicator score, index component (with values of Mobility, Workplace, etc.), and country group (which is represented by the colors of the bars in the graph and explained in the legend). The display also includes the average for all countries for each of the indicators, marked by the dashed lines and the value of the average.
- **What data distributions and relationships between variables are shown?** This graphical display highlights the variability in the overall averages for the eight different indicators and also the variability in the country group averages for each of the indicators. It is clear from the display that the differences in the averages for the country groups are not the same across all of the indicators.
- **What story does the graphical display tell?** There are several interesting things that are communicated by this display. Some country groups tend to score above average across the eight indicators, while other country groups tend to score below the average. The OECD high-income group of countries has the highest average score on six of the indicators, and it is equal to or nearly equal to the highest score on the other two indicators, suggesting that the OECD high-income countries tend to provide more economic opportunities for women than countries in the other country groups.

  The parenthood indicator received the lowest overall average score, and this was also the lowest-scoring indicator for all the country groups except for the Europe & Central Asia group. This is the conclusion that is referenced in the title of display: "Economies have the lowest scores on the parenthood indicator." The Middle East & North Africa country group tended to score lower than the other county groups on most of the indicators. There is certainly room for improvement for the countries in every country group on at least some of the indicators.

What Do You Think?

What is an example of another conclusion that you could draw from this graphical display?

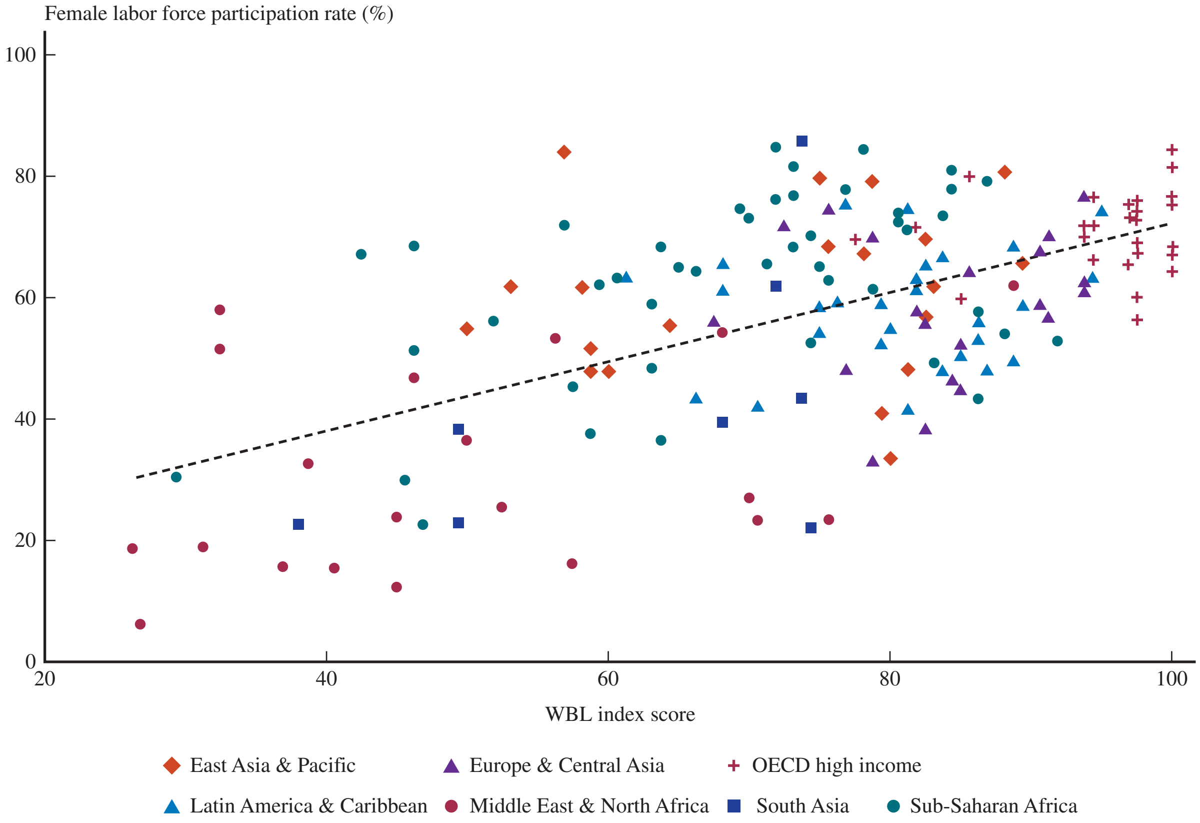
- **Is there room for improvement?** This graphical display does do a good job of communicating information about how the different country groups are doing on each of the indicators. The graph makes it easy to compare country groups for each indicator and to compare worldwide averages across indicators. Suggestions for improvement might include revising the title to be more descriptive of the entire graphical display and indicating somewhere in the display that the numerical values represented in the graph are averages.

Figure 3.46 is another graphical display that appears in the report.

There are three variables represented in the display in Figure 3.46: WBL index score, female labor participation rate (as a percentage), and country group (indicated by color and symbol type, as identified in the legend). You will recognize this as a scatterplot of WBL index score versus female labor participation rate. Each point in the plot represents one of the 176 countries, but colors and symbol shapes are used to distinguish the data points by country group.

Figure 3.46

Scatterplot of WBL Index score versus female labor force participation rate by country group.



Source: *Women, Business and the Law* (WBL) database; World Development Indicators (WDI) database, World Bank, https://databank.worldbank.org/source/world-development-indicators; International Labour Organization (ILO), https://ilostat.ilo.org/.

This graphical display allows you to see the relationship between WBI index score and female labor participation rate. It also allows you to see what that relationship looks like for each of the country groups and compare the relationship for different country groups, although this is a bit difficult to see. This would have been easier if the display had also included trend lines for each of the country groups. The line in the graphical display calls attention to the overall positive relationship, with countries that have greater WBL index scores tending to have greater female labor participation rates.

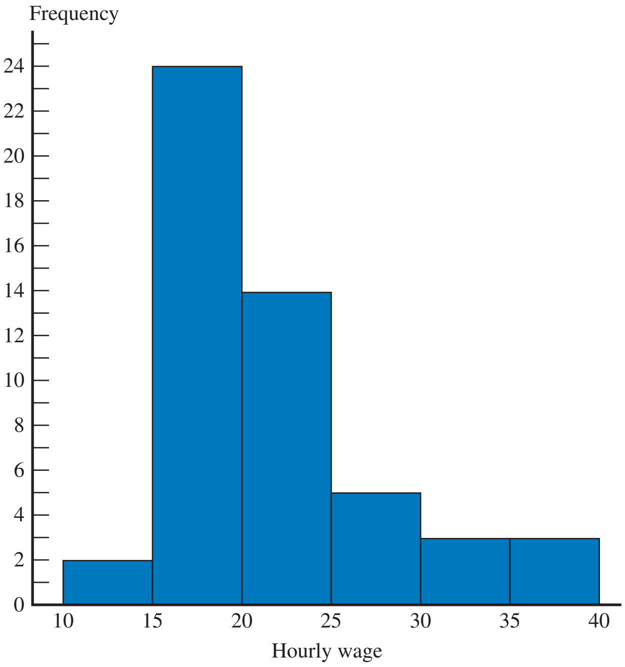
Example 3.28 The High Cost of Housing Revisited

Many people struggle to afford rental housing, and the cost of rental housing varies widely from state to state. Examples 3.15 and 3.16 used data from the report **"Out of Reach" (National Low Income Housing Coalition, 2021, nlihc.org)** on the hourly wage a full-time worker would need to earn in order to afford a two-bedroom rental home without spending more than 30% of their income on housing for each of the 50 U.S. states and the District of Columbia. In those examples, the data distribution was summarized in a frequency distribution and a histogram. The histogram appears again here in Figure 3.47. Based on the histogram, you can see that an hourly wage above $20 would be needed to afford a two-bedroom rental in about half the states and that the hourly wage needed for six states is over $30.

The report referenced also summarized these data using a graphical display similar to the one in Figure 3.48. Notice that this graphical display introduces the variable state and includes it by displaying the data on hourly wage on a map of the United States. The actual data values for each state are shown, and the states are also color-coded using categories described in the legend of the display. Presenting the data in this way provides insight into where the problem of affordable housing is the greatest. For example, you can see that for states on the West Coast and the East Coast, the hourly wages required for a full-time

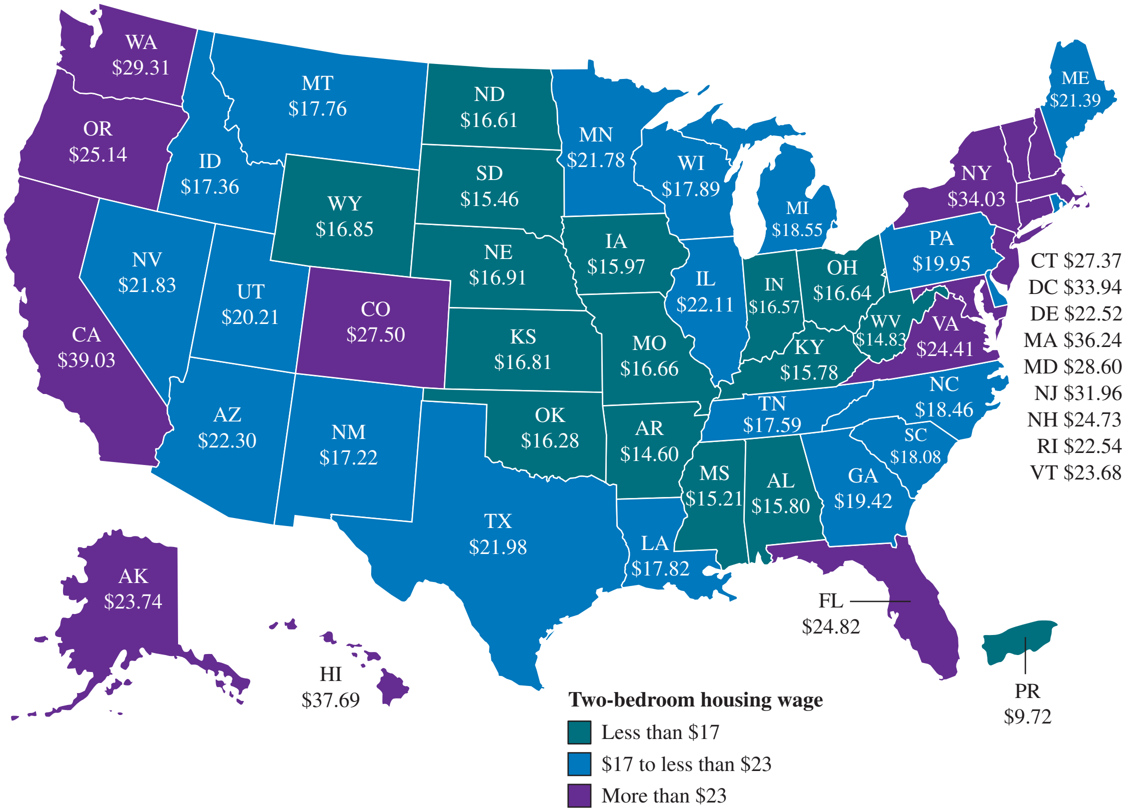
Copyright 2025 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s). Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

Figure 3.47
Histogram of the hourly wage data.



worker to afford a two-bedroom rental tend to be higher. This information is not communicated when the data are summarized in a histogram. However, it is difficult to see the actual data distribution for the hourly wage or to identify any potential outliers when the data is displayed in the form of a map. An interesting next step would be to explore the relationship between the hourly wage necessary to afford a two-bedroom rental home without spending more than 30% of income on housing and the actual minimum wage for the 50 states.

Figure 3.48
Hourly wage data displayed on a map.


WA $29.31
OR $25.14
CA $39.03
MT $17.76
ID $17.36
NV $21.83
ND $16.61
SD $15.46
WY $16.85
UT $20.21
CO $27.50
AZ $22.30
NM $17.22
MN $21.78
NE $16.91
KS $16.81
OK $16.28
TX $21.98
IA $15.97
MO $16.66
AR $14.60
LA $17.82
WI $17.89
IL $22.11
MS $15.21
MI $18.55
IN $16.57
KY $15.78
TN $17.59
AL $15.80
OH $16.64
WV $14.83
GA $19.42
FL $24.82
SC $18.08
NC $18.46
VA $24.41
PA $19.95
NY $34.03
ME $21.39
CT $27.37
DC $33.94
DE $22.52
MA $36.24
MD $28.60
NJ $31.96
NH $24.73
RI $22.54
VT $23.68
AK $23.74
HI $37.69
PR $9.72
Two-bedroom housing wage
Less than $17
$17 to less than $23
More than $23

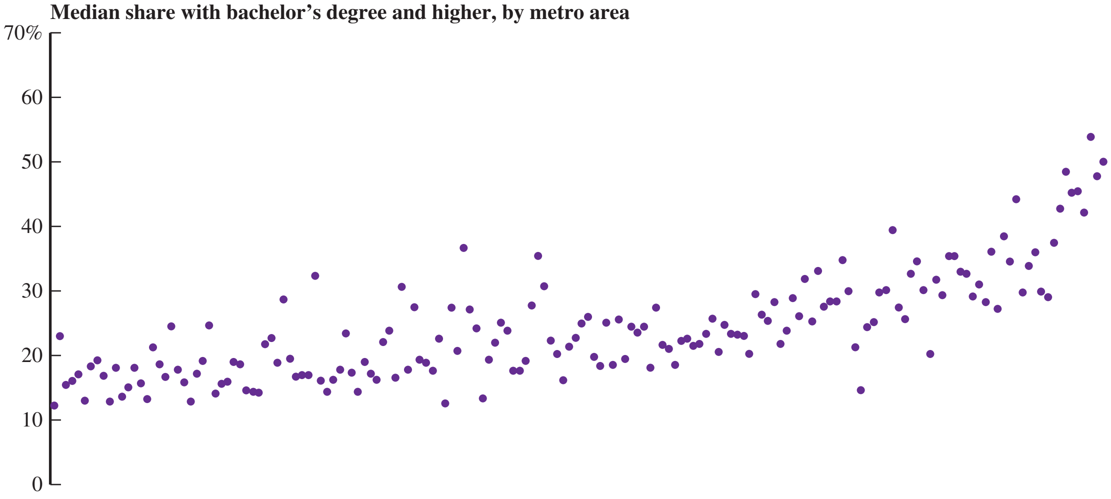
Example 3.29 Education Gaps

The article **"The Gaps Between White and Black America, in Charts" (*The New York Times*, June 19, 2020, nytimes.com/interactive/2020/06/19/opinion/politics/opportunity-gaps-race-inequality.html)** considers racial inequality across a number of socioeconomic and demographic variables, including household income, unemployment rate, percent of the population with a bachelor's degree, and home ownership rate. Several interactive graphs are included in the article. The graphical displays in this article were produced using data from the American Community Survey, the Centers for Disease Control and Prevention, the Census Bureau, and the Bureau of Labor Statistics. It is difficult to talk about an animated graph in print, so you are encouraged to visit the website yourself to get a real sense of the power of animated graphical displays.

The graph in Figure 3.49 is similar to what you first see in the online article. It is a display of the median percent of the population that has a bachelor's degree for communities in major metropolitan areas. Each dot in the graph represents one metropolitan area, and the height of the dot in the display represents the median value when the percentage with a bachelor's degree is recorded for each community in that metropolitan area. For example, the dot on the far right (corresponding to a value of about 45%) represents San Francisco and tells you that the median value for the communities in San Francisco is about 45%. In the interactive graph that appears on the website, you can move your cursor over any dot to display the name of the corresponding metropolitan area.

Figure 3.49

Graph of median percent with a bachelor's degree for different metropolitan areas.

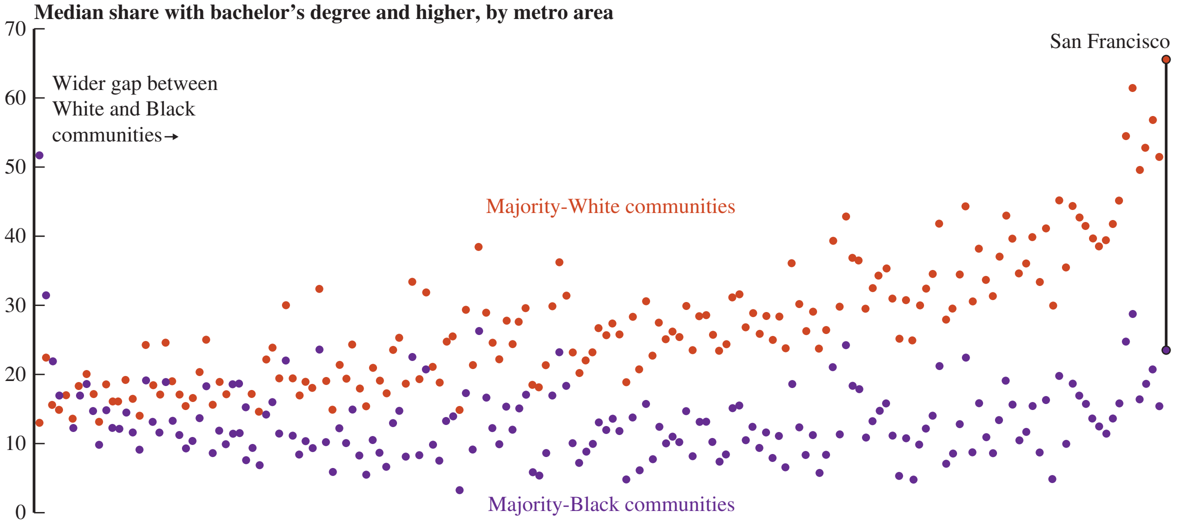


From this starting point, you can see that the percentage with a bachelor's degree varies quite a bit across metropolitan areas, with some having medians in the 10 to 20 percent range and others having medians that were greater than 30%. But this doesn't really tell the whole story. When you animate this display, it separates the communities in each metropolitan area into those where the majority of residents are White and those where the majority of residents are Black and then displays the medians for these two subgroups. Figure 3.50 is similar to what you see on the website after the animation. For each metropolitan area, the median for the majority-Black communities and the median for the majority-White communities are given. The line segment shown for San Francisco represents the gap between the percent who have a bachelor's degree in the two neighborhood types. The metropolitan areas are ordered from left to right based on the size of the gap.

Several things are revealed by this graph. There is a lot of variability in the size of the gap for different metropolitan areas, and there is more variability in the median percentages in the majority-White communities than in the majority-Black communities. Not surprisingly, the gap is smaller for metropolitan areas where the percentages tend to be low. In the interactive graph that appears on the website, you can move your cursor over the dots to see which metropolitan areas have relatively large gaps and which have smaller gaps.

Figure 3.50

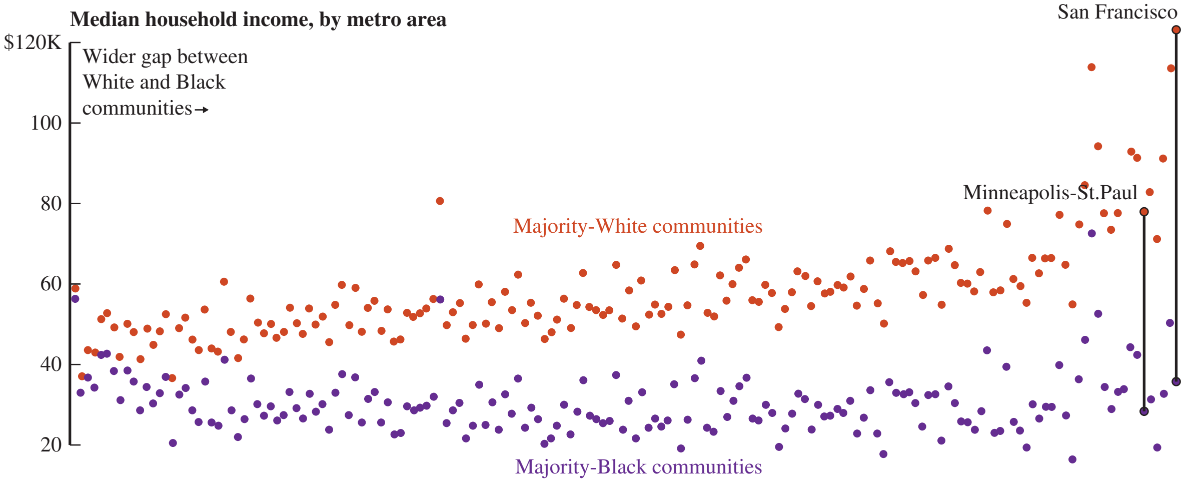
Graph of median percent with a bachelor's degree for different metropolitan areas by community type.



Similar graphs that illustrate the gap that exists in other variables are also included on the website. For example, Figure 3.51 shows a similar graph that focuses on household income.

Figure 3.51

Graph of median household income for different metropolitan areas by community type.



What Do You Think?

What do you notice about the relationship between community type and median household income for the metropolitan areas represented in Figure 3.51?

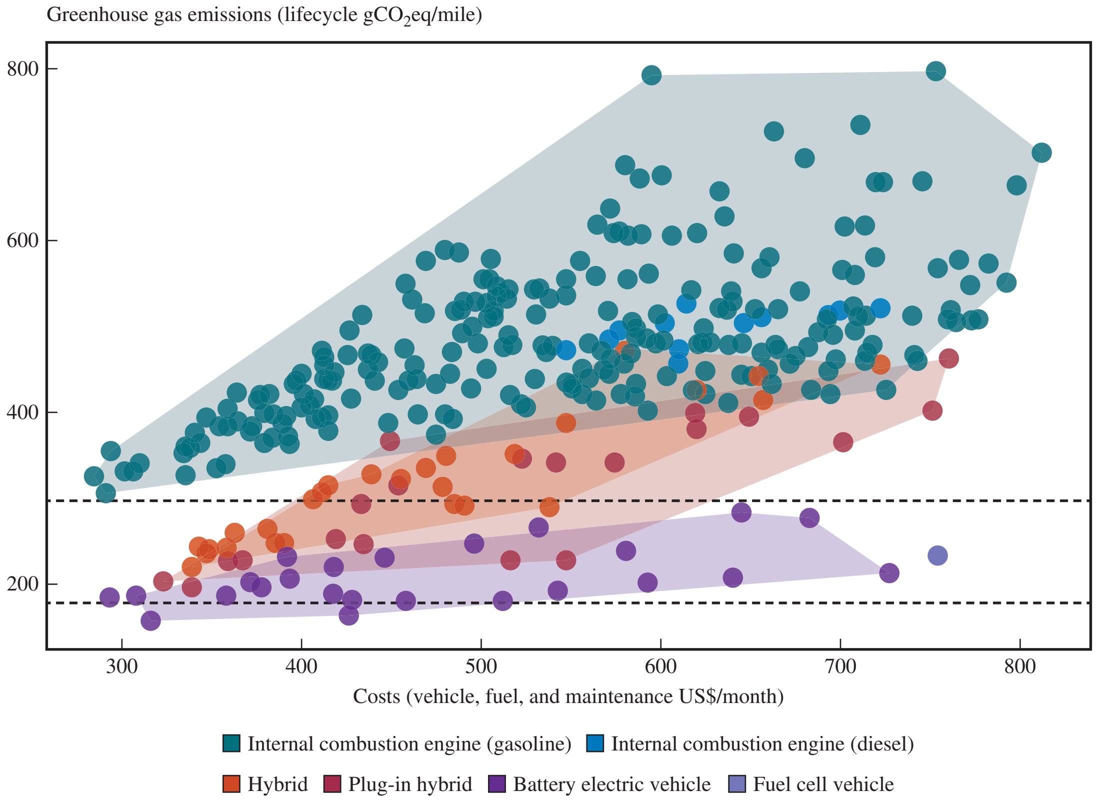
In what ways is the relationship between community type and median household income different from the relationship between community type and percentage with a bachelor's degree that is illustrated in Figure 3.50?

Example 3.30 Greenhouse Gas Emissions

The graphical display in Figure 3.52 is similar to one that appears on the Carbon-Counter website (**carboncounter.com**). This graph shows the average cost per month (in dollars) and the greenhouse gas (GHG) emissions per mile for cars and SUV models manufactured in 2021. The purchase price of the vehicle, typical maintenance costs, and cost of fuel (gas or electricity) based on typical driving mileage per year were all used to determine the average cost per month. The GHG emissions for a particular vehicle were based on the emissions associated with vehicle use, as well as emissions associated with vehicle manufacturing and with fuel or electricity production. More information on how these variables were determined can be found in the paper **"Personal Vehicle Emissions Evaluated Against Climate Change Mitigation Targets" (pubs.acs.org)**.

Figure 3.52

Graphical display of vehicle emissions, cost, and type of engine for 2021 vehicles.



In addition to the two variables already mentioned, notice that the graphical display also used color to provide information on a third variable—vehicle type. The designers of this graph have also shaded regions that capture the vehicles of a given type to help communicate differences between the types of vehicles.

What Do You Think?

Many people believe that cars with lower greenhouse emissions cost more than those with higher greenhouse gas emissions. Is that belief supported by the data used to construct this graphical display?

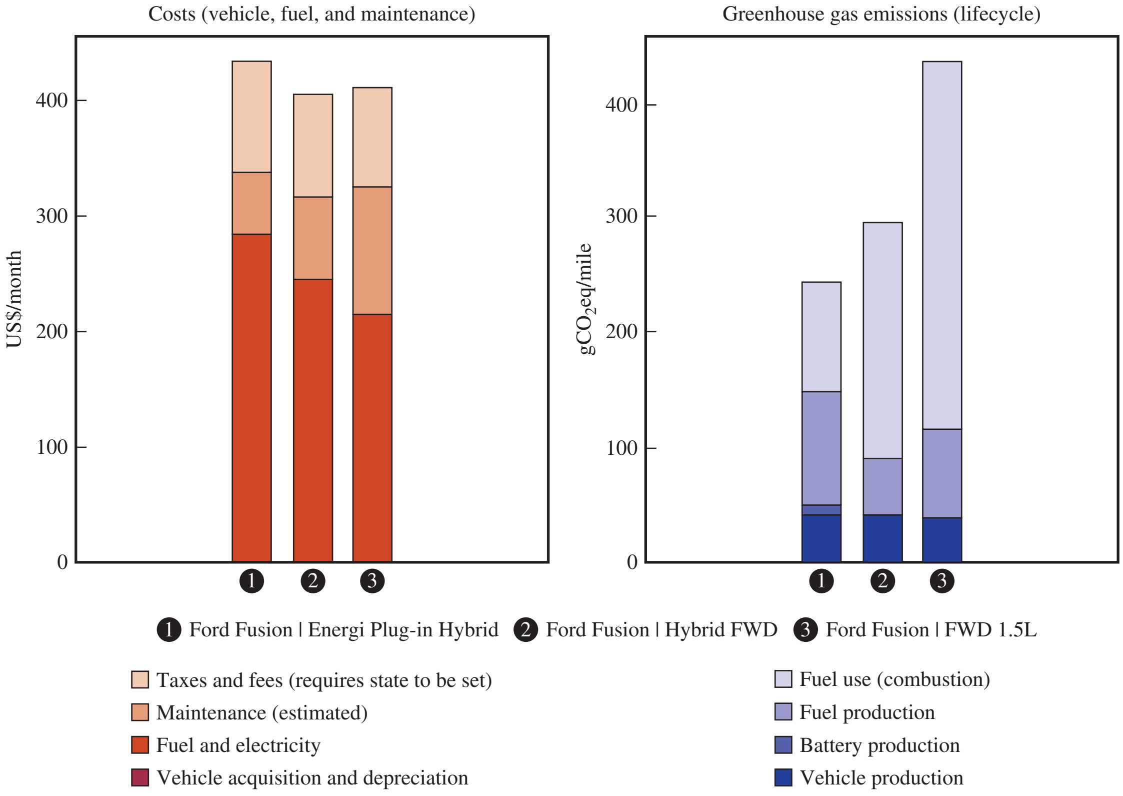
There are several interesting messages communicated in this graphical display. It does appear that the relationship between cost and emissions is different for different types of vehicles. For battery electric vehicles, there does not appear to be much of a relationship between cost and emissions, but for internal combustion vehicles, the emissions do tend to increase as the cost of the vehicle increases. The paper referenced earlier includes the following statement, which uses these abbreviations: HEVs for hybrid, PHEVs for plug-in hybrid, BEVs for battery electric, and ICEV for internal combustion.

> We find that GHG emissions and costs vary considerably across popular vehicle models, both within and across powertrain technologies, with lower emissions generally corresponding to lower costs. Alternative powertrain technologies (HEVs, PHEVs, and BEVs) exhibit systematically lower lifecycle GHG emissions than ICEVs but do not necessarily cost the consumer more.

Another interesting aspect of this graphical display is the two dashed lines. You might be wondering what those represent. The upper line represents the U.S. emission-reduction target for the year 2030, and the lower line represents the target for the year 2040. Notice that all the electric vehicles and many of the hybrid and plug-in hybrid vehicles already meet the 2030 target, while none of the internal combustion vehicles currently meets that target.

The graphical display on this website is also interactive. If you hover over a particular point in the display, it will show details for that vehicle. It also lets you select particular vehicles for further analysis. For example, the Ford Fusion comes in an internal combustion model, a hybrid model, and a plug-in hybrid model. Selecting these three vehicles and then clicking on the barchart tab produces a graphical display similar to the one shown in Figure 3.53. This allows you to compare the cost and emissions across the three vehicles. Notice that the overall cost of the three vehicles is similar, but that there are large differences in emissions. Emissions are lowest for the plug-in hybrid model, even after taking into account the emissions involved with manufacturing the batteries that the car uses and the production of the electricity and gas that the vehicle uses.

Figure 3.53
Comparing three Ford Fusion Models.



The graphical displays you have seen in the examples of this section are just a few of the many different types of displays you might encounter. Some additional examples can be found in the exercises, but there are lots of interesting displays and infographics being produced on a daily basis and you are sure to see them in newspapers, magazines, journals, and online.

Exercises 3.48–3.60

3.48 The graphical display at the bottom of the page is similar to one that appeared in the article **"The Pandemic Changed How We Spent Our Time" (*New York Times*, July 27, 2021)**. The display was constructed using data from the American Time Use Survey, which asks participants to provide a detailed account of how they spent their time during one 24-hour period.

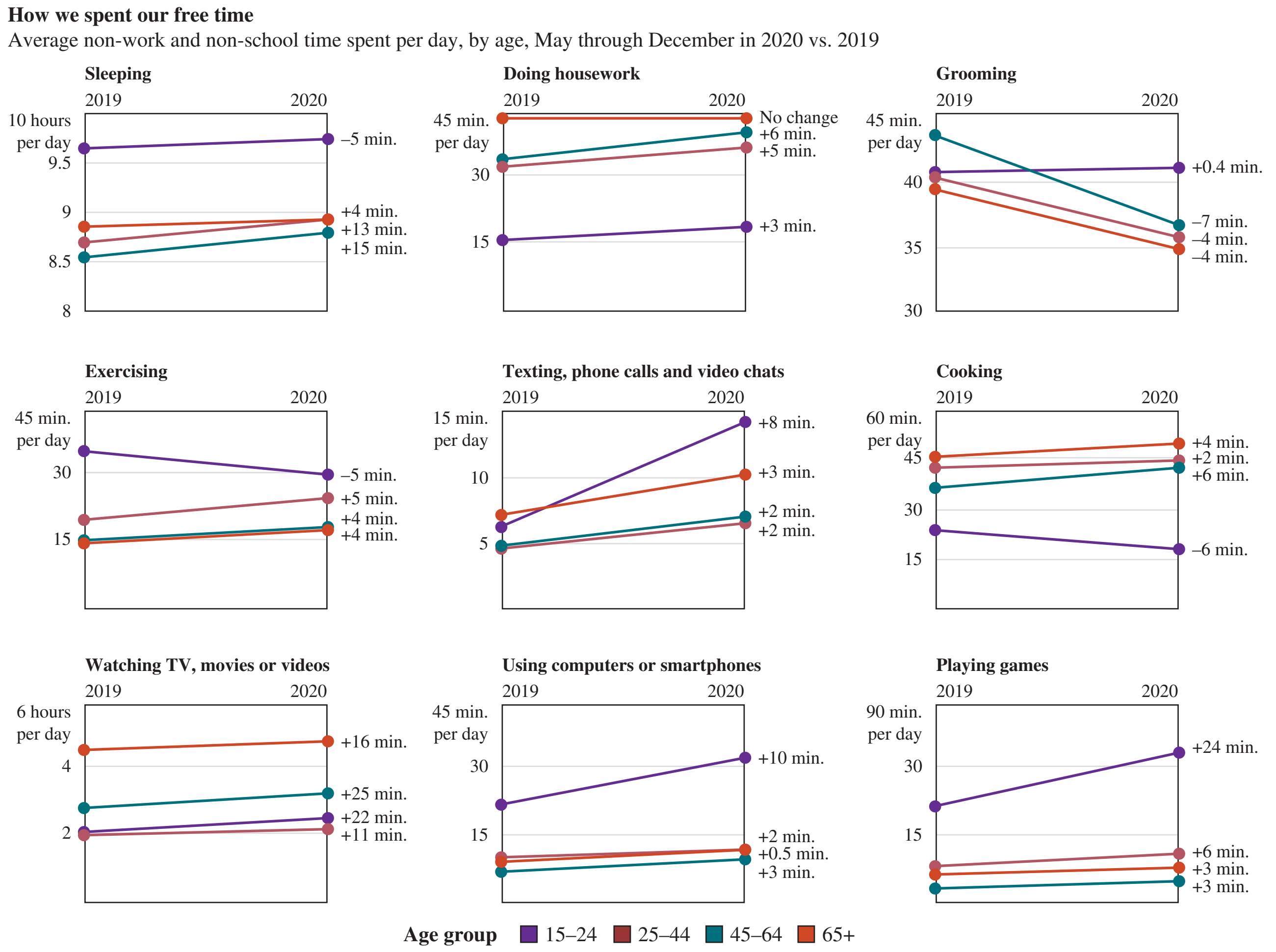
- **a.** There are four variables represented in this graphical display. What are they?
- **b.** Did time spent on grooming tend to increase or decrease between 2019 and 2020? How does the change from 2019 to 2020 depend on age group?
- **c.** Does any one age group stand out as unusual when you look at this display? If so, describe how this age group tends to be different from the other age groups?
- **d.** Write a brief summary of the story that this graphical display tells.
- **e.** Does the display do a good job in communicating its message?

3.49 A histogram of the median age for each of the 50 states and the District of Columbia is shown at the top of the next page.

- **a.** Comment on any interesting features of the histogram and give a brief description of the data distribution.
- **b.** A graphical display similar to one that appeared in the article **"Here's How Old the Typical Person is in Each State" (businessinsider.com)** is also shown. What can you learn from this graphical display that you can't see in the histogram? What can you learn from the histogram that is difficult to see in the map display?

3.50 Do you enjoy looking at the night sky? What you can see depends a lot on the level of light pollution in the area. The National Park Service defines light pollution as the introduction of artificial light into the natural environment. The image shown at the top of page 144 is from the National Park Service Night Skies webpage (**nps.gov**).

Figure for Exercise 3.48



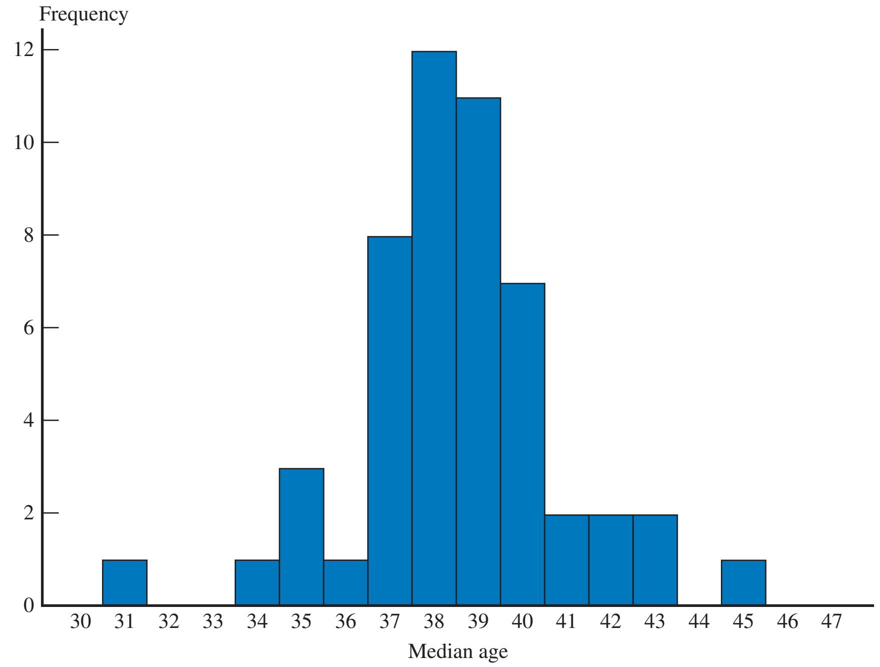
How we spent our free time

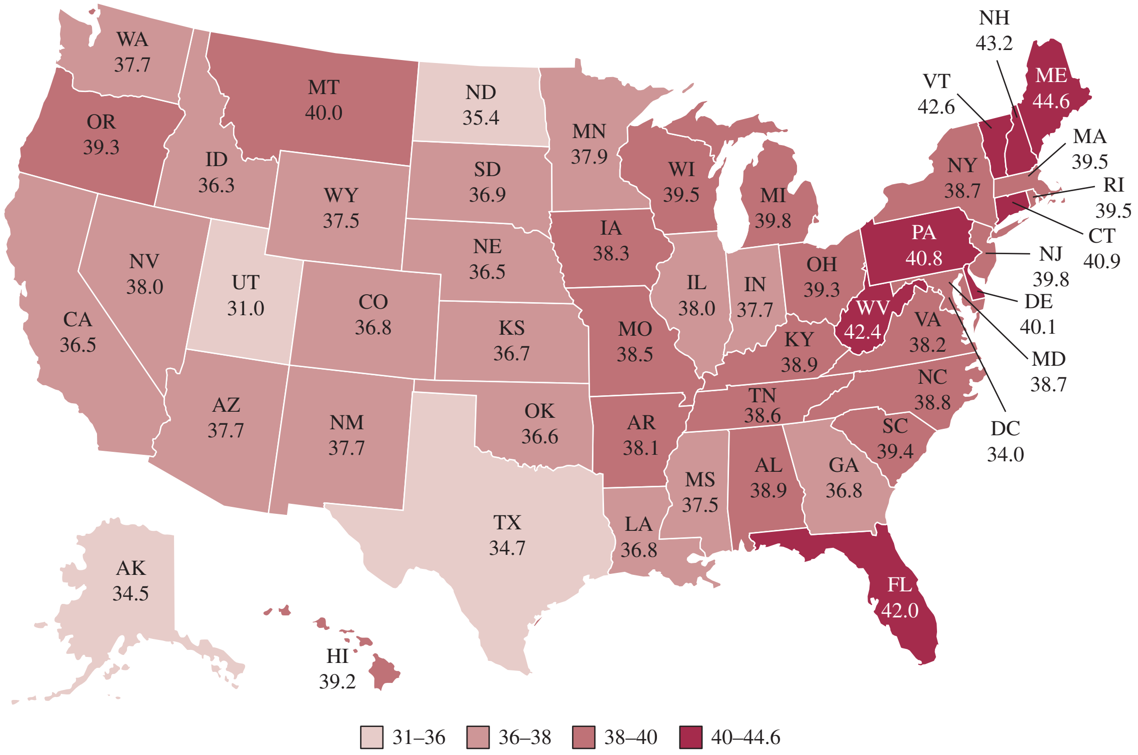
Average non-work and non-school time spent per day, by age, May through December in 2020 vs. 2019

Notes: Annual averages include all people, not just those who participated at all in each activity. The computers and smartphones category excludes school and work use, or time spent on devices for gaming, watching videos or other uses covered in other categories. Examples of included uses are social media and internet activity.

Copyright 2025 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s). Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

Figure for Exercise 3.49




WA 37.7
OR 39.3
CA 36.5
ID 36.3
NV 38.0
MT 40.0
WY 37.5
UT 31.0
AZ 37.7
CO 36.8
NM 37.7
ND 35.4
SD 36.9
NE 36.5
KS 36.7
OK 36.6
TX 34.7
MN 37.9
IA 38.3
MO 38.5
AR 38.1
LA 36.8
WI 39.5
IL 38.0
MS 37.5
MI 39.8
IN 37.7
KY 38.9
TN 38.6
AL 38.9
OH 39.3
WV 42.4
GA 36.8
FL 42.0
SC 39.4
NC 38.8
VA 38.2
PA 40.8
NY 38.7
VT 42.6
NH 43.2
ME 44.6
MA 39.5
RI 39.5
CT 40.9
NJ 39.8
DE 40.1
MD 38.7
DC 34.0
AK 34.5
HI 39.2
31–36
36–38
38–40
40–44.6

Source: US Census Bureau, 2017 American Community Survey

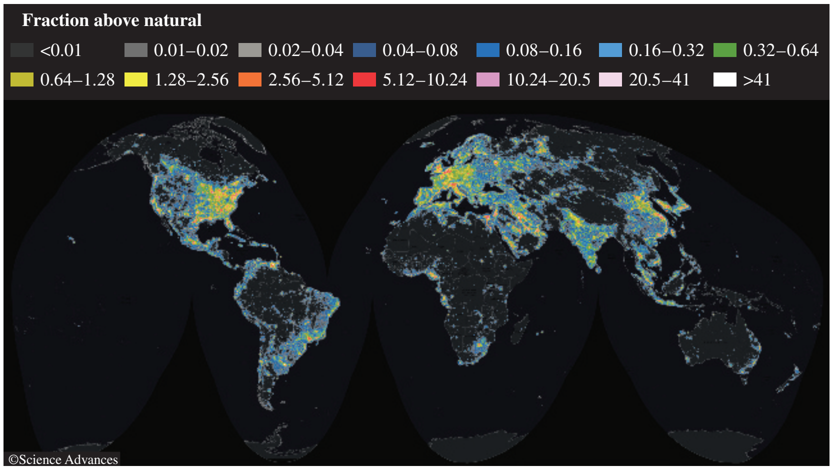
Exercise 3.50 continued

a. What variables are represented in this display? What can you learn about light pollution from this display?

b. Next go to this website: **cires.colorado.edu/artificial-sky**. Here you can get a closer look at light pollution in the United States. Click on "View Map Legend" (which can be found just below the map) to open a legend that explains the colors in the map. Then write a few sentences explaining what this display reveals about light pollution in the United States If you live in the United States, how does where you live compare to other parts of the United States in terms of light pollution?

Copyright 2025 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s). Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

Figure for Exercise 3.50



New World Atlas map shows extent of light pollution across the globe.

© Hinari, Agora, Oare, Chorus, Clockss, Crossref And Counter/Science Advances Eissn Fabio Falchi , Pierantonio Cinzano, Dan Duriscoe, Christopher C. M. Kyba, Christopher D. Elvidge, Kimberly Baugh, Boris A. Portnov, Nataliya A. Rybnikova, And Riccardo Furgon/AAAS

Exercise 3.50 continued

- **c.** If you wanted to get the best view of the night sky, where in the United States would you go, and why?
- **d.** Now click on "View 3D Globe version of this map" (this appears just above the map for the United States). This opens up an interactive display that allows you to navigate the globe and explore light pollution around the world. Spend a few minutes interacting with this display and then write a short paragraph describing what this display reveals about light pollution around the world.

3.51 The website **The Climate Lab Book (climate-lab-book.ac.uk/spirals/)** uses an interesting animated graphical display to illustrate climate change over time. There are three animated graphical displays on the website to track global temperature change, arctic sea ice volume, and atmospheric CO_2 concentration. Go to this website and select one of these displays. Click on the graph to start the animation. After watching the animation, answer the following questions.

- **a.** Why do you think the animation uses a spiraling motion to represent the variable of interest?
- **b.** What does one revolution around the spiral represent?
- **c.** What do you think the colors in the animation represent?
- **d.** What story does this animated graphical display tell?
- **e.** Does the display do a good job in communicating its message?

3.52 The report **"Women, Business and the Law 2021" (worldbank.org)** has several other interesting multivariable graphical displays in addition to the two discussed in Example 3.27, including the one shown at the top of the next page.

- **a.** What variables are represented in this graph?
- **b.** What do you notice about the relationships between the variables shown in this graph?
- **c.** What story does this graphical display tell?

3.53 The graphical display at the bottom of the next page is similar to one appearing in the report **"College Enrollment and Retention in the Era of Covid" (The College Board, June 2021, research.collegeboard.org)**.

- **a.** What variables are represented in this graphical display?
- **b.** What do the colors used in the display represent?
- **c.** Write a brief summary of the story that this graphical display tells.
- **d.** Does the display do a good job in communicating its message?

3.54 The report referenced in the previous exercise also includes a graphical display similar to the one shown on page 146.

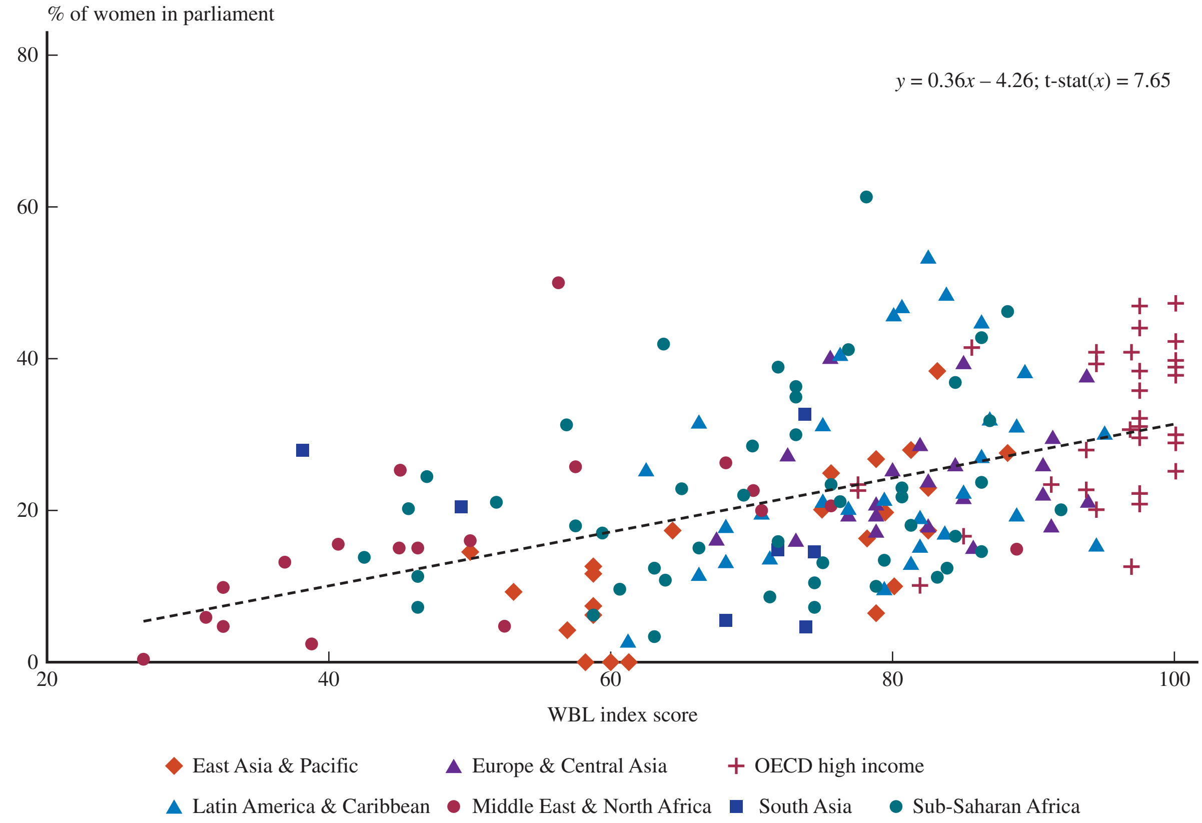
- **a.** What does this display reveal about the relationships between enrollment change, geographical region, and type of institution (four-year and public two-year)?
- **b.** Write a brief summary of the story that this graphical display tells.

3.55 Scientists with the National Park Service made long-term sound measurements in parks as well as in rural and urban areas around the United States. They used these data to produce the display of sound levels that appears at the top of page 147 **(nps.gov)**.

- **a.** What variables are represented in this display? What can you learn about noise levels from this display?
- **b.** You can find an interactive version of this display at this website: **nps.gov/subjects/sound/upload/CONUS_Existing_L50dBA_SummerDay_Legend.png**. There you can zoom in to get a closer look at a specific region or city. If you live in the United States, click to get a closer look at your region. How does where you live compare to other parts of the United States in terms of noise level? If you do not live in the United States, choose a region that is of interest to you.

Copyright 2025 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s). Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

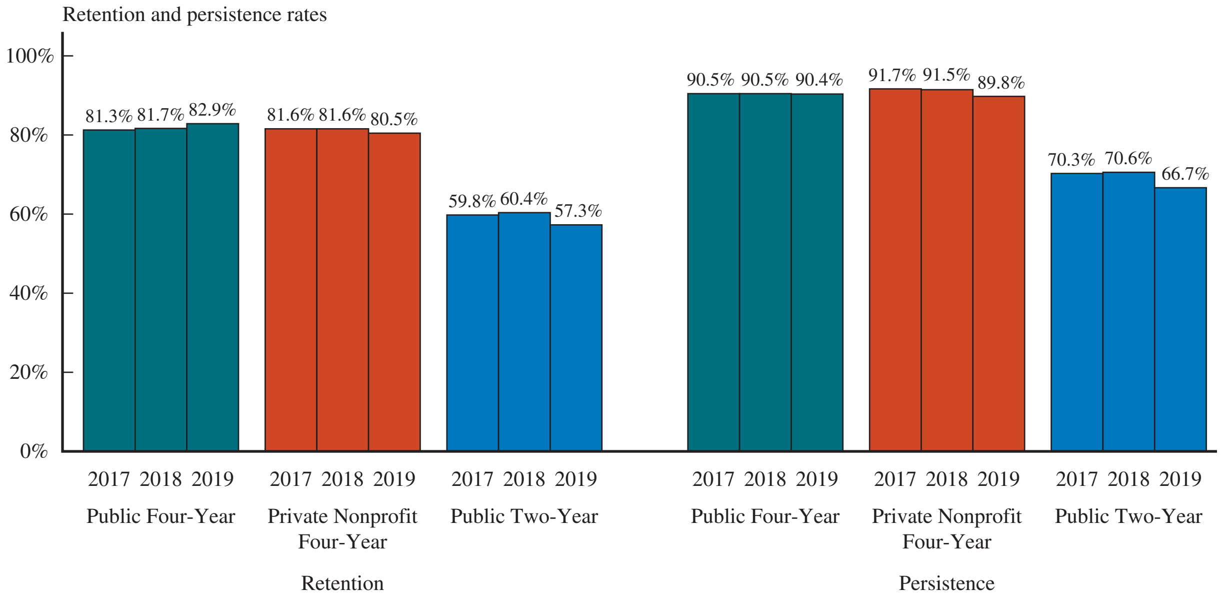
Figure for Exercise 3.52



Sources: *Women, Business and the law* (WBL) database; Inter-parliamentary Union (IPU)

Note: The Figure presents the correlation between the WBL index score and the percentage of parliamentary seats held by women based on data for 2019. Each point represents a single economy. A fitted regression line is also included. The relationship between the WBL index and the proportion of seats held by the women in parliament remains positive and statistically significant after controlling for income, measure as gross domestic product (GDP) per capita provided in the World Bank's world department indicators (WDI) database (https//databankworldbank.org/source/world-development-indicators), and economy-level and time fixed effects. This statistical relationship should not be interpreted as causal. The regression of the proportion of seat held by the women in parliament of the WBL index includes 183 economics for the period 1997–2019, as determined by data availability in OECD (Organisation for Economic Co-operation and Development).

Figure for Exercise 3.53



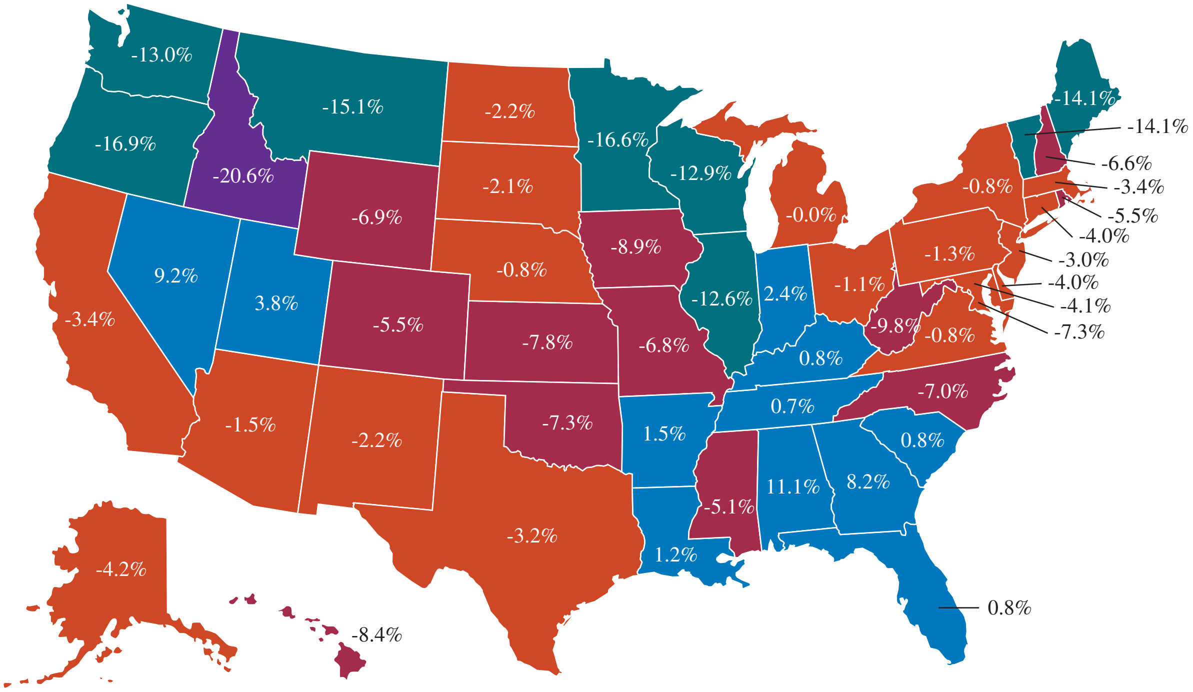
Note: First-year retention rates measure the share of students enrolling in the same college 1 year after initial enrollment. First-year persistence rates measure the share of students enrolling in any college 1 year after initial enrollment.

Copyright 2025 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s). Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

Figure for Exercise 3.54

Change in First-Time Student Enrollment from 2019 to 2020 by Student State of Residence

Four-Year


-13.0%
-15.1%
-2.2%
-16.6%
-14.1%
-14.1%
-16.9%
-20.6%
-2.1%
-12.9%
-6.6%
-0.8%
-3.4%
-6.9%
-0.0%
-5.5%
-4.0%
9.2%
-8.9%
-1.3%
-3.0%
-0.8%
-4.0%
3.8%
-12.6%
2.4%
-1.1%
-4.1%
-3.4%
-5.5%
-7.8%
-6.8%
-9.8%
-0.8%
-7.3%
0.8%
-7.0%
0.7%
-1.5%
-2.2%
-7.3%
1.5%
0.8%
11.1%
8.2%
-5.1%
-3.2%
1.2%
-4.2%
0.8%
-8.4%

Public Two-Year

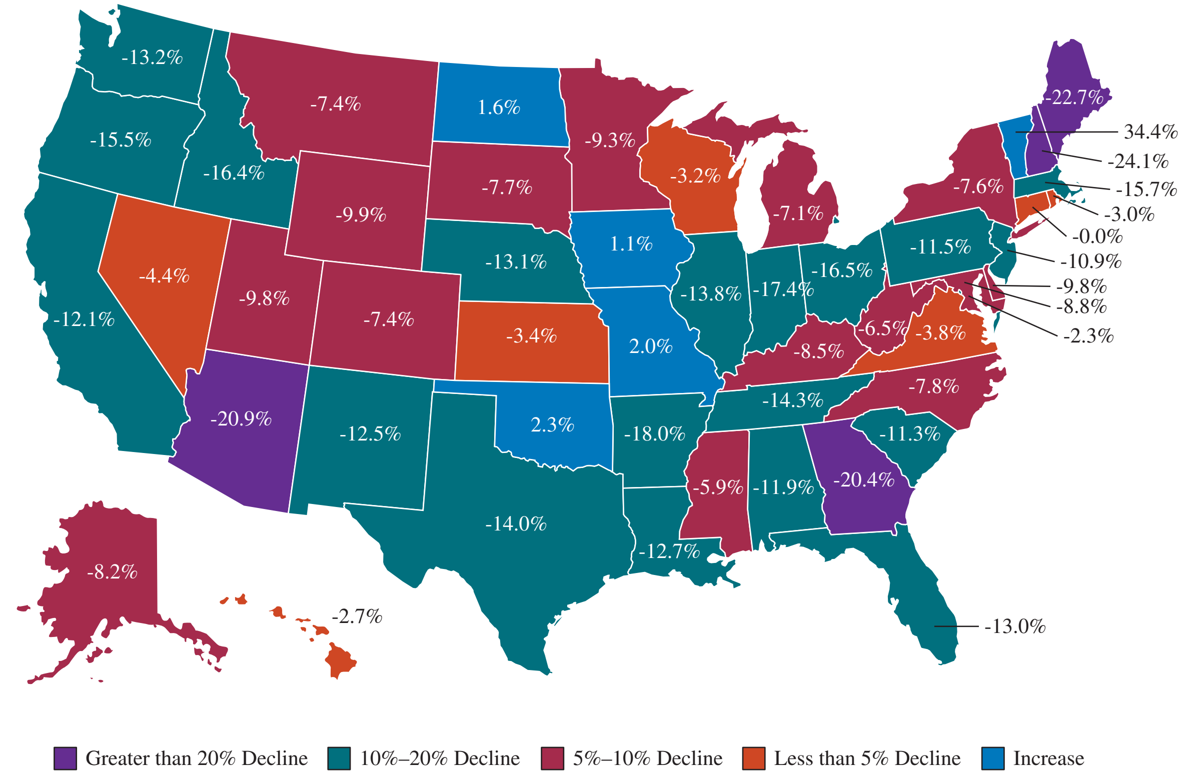
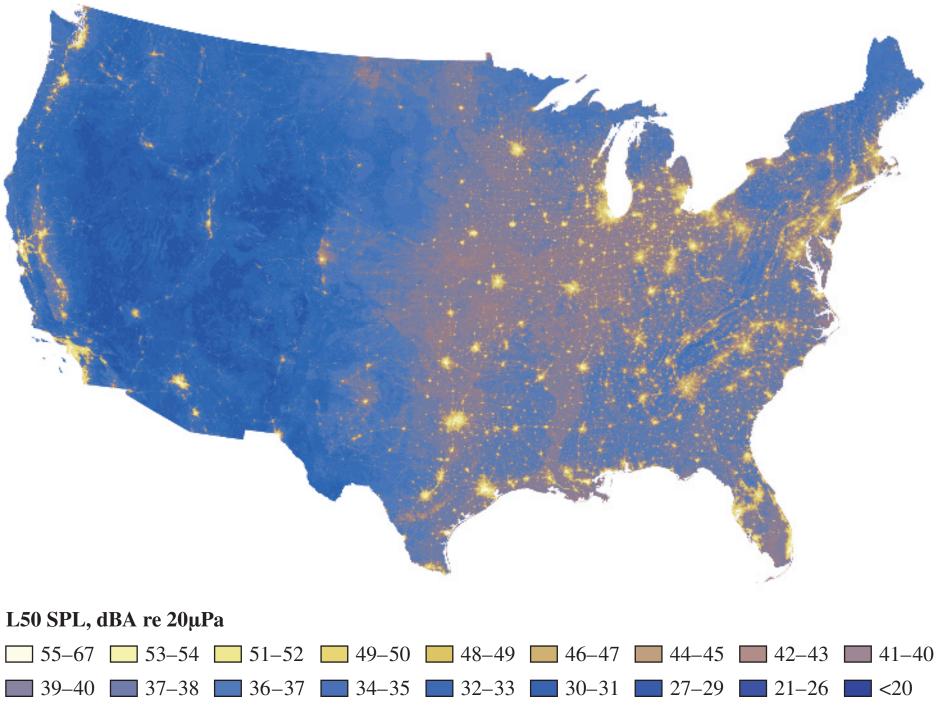

-13.2%
-7.4%
1.6%
-22.7%
34.4%
-15.5%
-9.3%
-24.1%
-16.4%
-7.7%
-3.2%
-7.6%
-15.7%
-9.9%
-7.1%
-3.0%
-0.0%
-4.4%
-13.1%
1.1%
-11.5%
-10.9%
-16.5%
-9.8%
-8.8%
-12.1%
-9.8%
-7.4%
-13.8%
-17.4%
-3.4%
2.0%
-6.5%
-3.8%
-2.3%
-8.5%
-7.8%
-14.3%
-20.9%
2.3%
-12.5%
-18.0%
-11.3%
-5.9%
-11.9%
-20.4%
-14.0%
-12.7%
-8.2%
-2.7%
-13.0%
Greater than 20% Decline
10%–20% Decline
5%–10% Decline
Less than 5% Decline
Increase

Figure for Exercise 3.55



3.56 The report **"Women, Business and the Law 2021" (worldbank.org)** has several other interesting multivariable graphical displays in addition to the two discussed in Example 3.27, including the one shown at the bottom of the page.

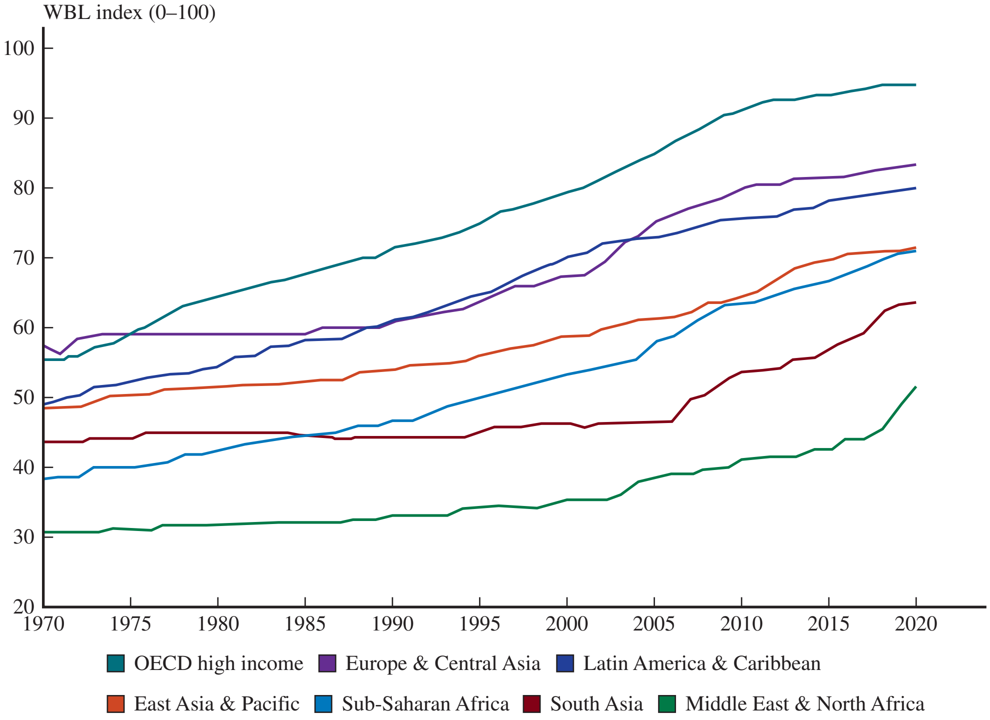
a. What variables are represented in this graph?

b. What do you notice about the relationships between the variables in this graph?

c. What story does this graphical display tell?

3.57 **"What Is Going On in This Graph"** is a feature of the ***New York Times* Learning Network**, where graphical displays are presented each week as a part of an online discussion. An archive of all of the graphs that have been featured can be found at **nytimes.com/column/whats-going-on-in-this-graph**. Visit this website and scroll down to view the featured graphs. Select a graph that is of interest to you and that includes information on three or more different

Figure for Exercise 3.56



Source: *women Business and the law* (WBL) database.

Note: Figure shows change in the average WBL index score across seven regions between 1970 and 2020 = organisation for Economic Co-operation and Development.

Copyright 2025 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s). Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

variables. Identify the variables represented in the graphical display and then write a few sentences describing the story that the graphical display communicates.

3.58 A graphical display similar to the one shown at the bottom of the page appears in the report **"Early Evidence of Medicaid's Important Role in School-based Health Services" (childtrends.org, December 14, 2020)**.

a. What variables are represented in this graphical display?

b. What do the colors used in the display represent?

c. What do the two shapes used in the display represent?

d. What does the size of the shape represent?

e. This display is not quite as accurate as a map of the United States, but the states and the District of Columbia do appear in the correct general region of the country. What does this display tell you about the relationship between Medicaid expenditure and geographic location?

f. Write a brief summary of the story that this graphical display tells.

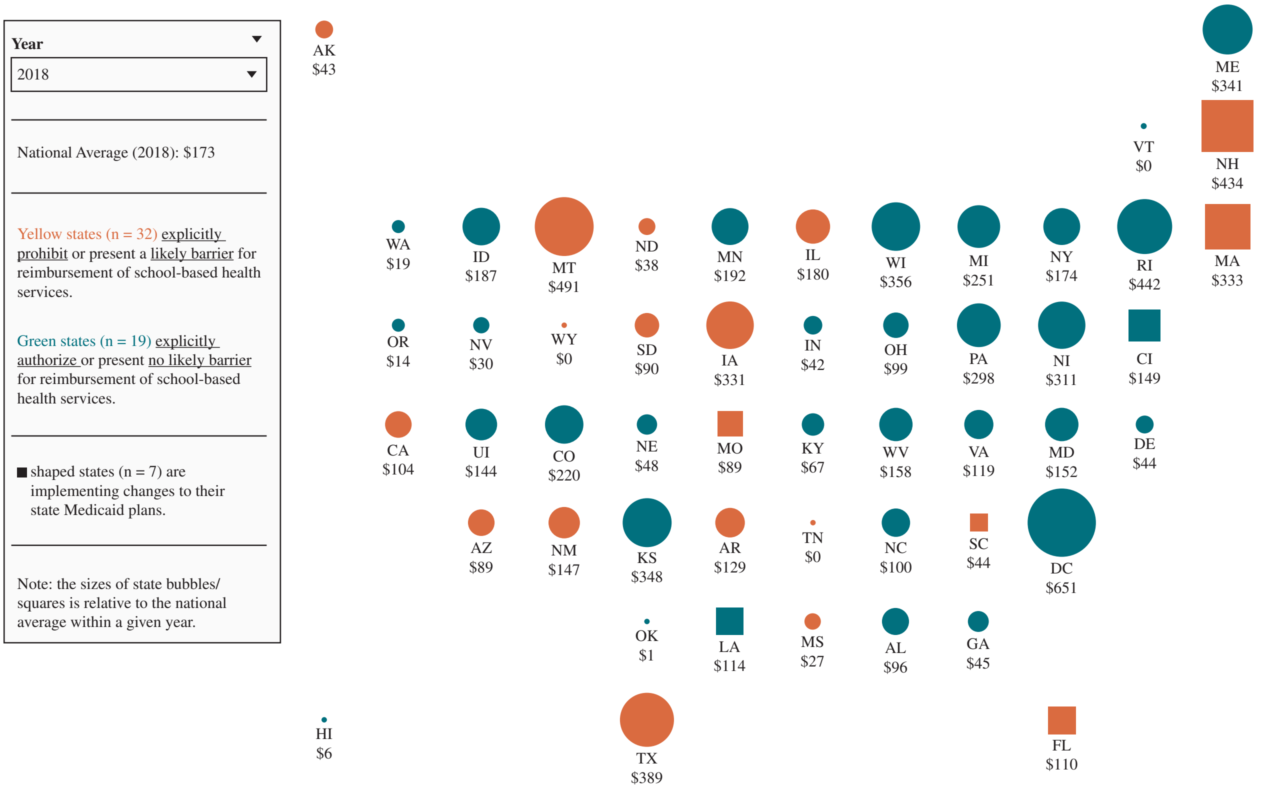
g. The data used to create this display could have been presented in a table that lists the states in one column and the Medicaid expenditure for each state in a second column. Would that have been a better way to communicate the intended message? Explain the thinking that led to your answer.

3.59 Kiva (**kiva.org**) is a nonprofit organization that uses crowdfunding to provide small loans to people around the world. The website **hempfieldkiva.org/data-visualization-of-six-years-of-kiva-loans/** has an interesting animated display based on data collected on Kiva loans. Go to this website and watch the short video. Then write a short summary of what you learned about Kiva loans from this animated graphical display.

3.60 Refer back to Example 3.30 for a discussion of the graphical displays contained on the **CarbonCounter** website (**carboncounter.com/#!/explore**). Follow the instructions on how you can interact with the graphical displays to select 3 different vehicles and produce a display similar to the one in Figure 3.53. Once you have produced the display, describe what is revealed about the relationships between cost, emissions, and vehicle model for the vehicles you selected.

Figure for Exercise 3.58

State Medicaid Expenditures on School-Based Health Services per Enrolled Child, by Potential Policy Barriers



Year ▼

2018 ▼

National Average (2018): $173

Yellow states (n = 32) explicitly prohibit or present a likely barrier for reimbursement of school-based health services.

Green states (n = 19) explicitly authorize or present no likely barrier for reimbursement of school-based health services.

■ shaped states (n = 7) are implementing changes to their state Medicaid plans.

Note: the sizes of state bubbles/squares is relative to the national average within a given year.

Copyright 2025 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s). Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

Section 3.6 Interpreting and Communicating the Results of Statistical Analyses

A graphical display, when used appropriately, can be a powerful tool for organizing and summarizing data. By sacrificing some of the detail of a complete listing of a data set, important features of the data distribution are more easily seen and more easily communicated to others.

Communicating the Results of Statistical Analyses

When reporting the results of a data analysis, a good place to start is with a graphical display of the data. A well-constructed graphical display is often the best way to highlight the essential characteristics of the data distribution, such as shape and variability for numerical data sets or the nature of the relationship between the two variables in a bivariate numerical data set.

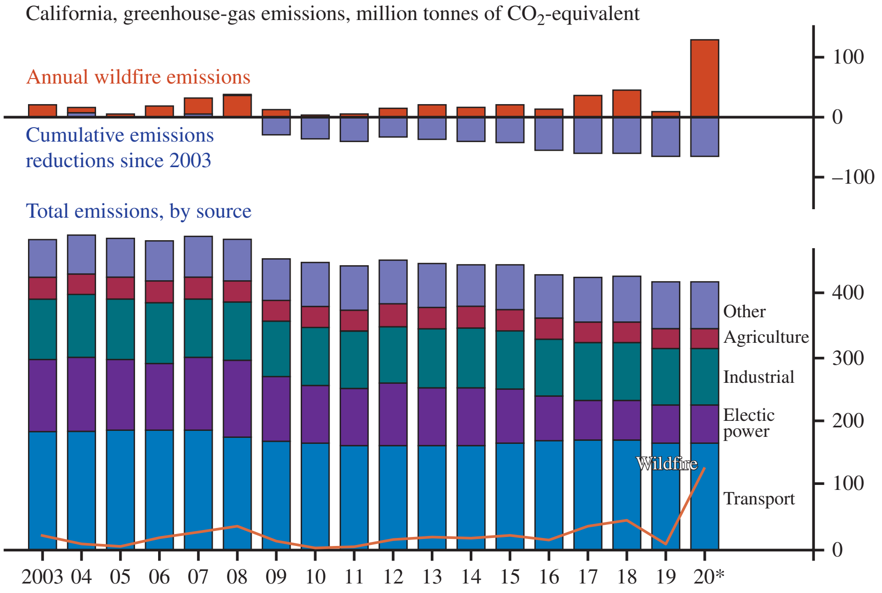
For effective communication with graphical displays, some things to remember are

- Be sure to select a display that is appropriate for the given type of data.
- Be sure to include scales and labels on the axes of graphical displays.
- In comparative plots, be sure to include labels or a legend so that it is clear which parts of the display correspond to which samples or groups in the data set.
- Although it is sometimes a good idea to have axes that do not cross at (0, 0) in a scatterplot, the vertical axis in a bar chart or a histogram should always start at 0 (see the cautions and limitations later in this section for more about this).
- Keep your graphs simple. A simple graphical display is much more effective than one that has a lot of extra "junk." Most people will not spend a great deal of time studying a graphical display, so its message should be clear and straightforward.
- Keep your graphical displays honest. People tend to look quickly at graphical displays, so it is important that a graph's first impression is an accurate and honest portrayal of the data distribution. In addition to the graphical display itself, reports often include a brief discussion of the features of the data distribution based on the graphical display.
- For categorical data, the discussion of the graphical display might be a few sentences about the relative proportion for each category, possibly pointing out categories that were either common or rare compared to other categories.
- For numerical data sets, the discussion of the graphical display usually summarizes the information that the display provides on three characteristics of the data distribution: center or location, variability, and shape.
- For bivariate numerical data, the discussion of the scatterplot typically focuses on the nature of the relationship between the two variables used to construct the plot.
- For data collected over time, any trends or patterns in the time series plot should be described.

Interpreting the Results of Statistical Analyses

There are many sources of greenhouse-gas emissions, including gas-powered cars and trucks, agriculture, and manufacturing. But emissions created by wildfires has also become a major contributing factor. The article **"One Year of Wildfires Undid Decades of California's Emissions Policy" (*The Economist*, October 19, 2022)** notes that even though some of the carbon emissions from wildfires will be reabsorbed as vegetation regrows, that process will take decades. They reach the conclusion that it makes sense to take wildfire emissions into account when considering greenhouse gases and the resulting effect on climate change. The authors of the article supported their conclusions with a graphical display similar to the one in Figure 3.54.

Figure 3.54 California greenhouse-gas emissions.



You can learn a great deal from this graphical display. The bottom graph in Figure 3.54 combines aspects of a time series plot with a segmented bar chart. It is designed to show both change over time as well as the changes in the sources of greenhouse gases for each year. It is interesting to note that in California, the emissions attributable to transport has remained about the same over time, and that the biggest decreases over time have been in the generation of electric power. The red line that has been added to the graph shows what contributions from wildfires were in each of the years. This source was not included in the segmented bar charts.

What Do You Think?

What would be the impact of adding in the emissions due to wildfires? How is this related to the title of the article?

The top graph of Figure 3.54 shows both the decrease in the greenhouse-gas emissions over time when wildfires are not considered and the increased contribution of wildfires over time. The two graphs in Figure 3.54 support the authors conclusion that wildfires "have wiped out the benefits of more than a decade of efforts to reduce emissions." Even so, without the progress on reducing emissions from other sources, the situation would be even more dire.

The use of graphical data displays is quite common in newspapers, magazines, and journals. For example, data on test scores for a standardized math test given to eighth graders in 37 states, 2 territories (Guam and the Virgin Islands), and the District of Columbia were used to construct the stem-and-leaf display and histogram shown in Figure 3.55. Careful examination of these displays reveals the following:

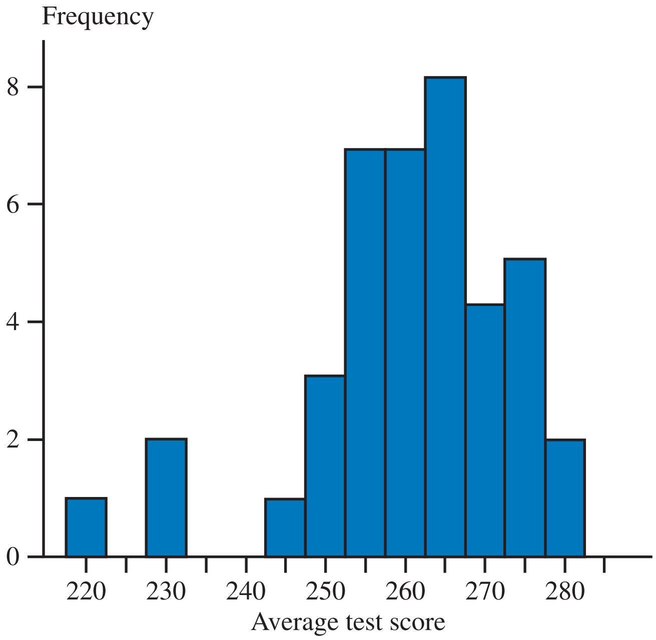
1. Most of the participating states had average eighth-grade math scores between 240 and 280. The shape of this display would be described as negatively skewed, because of the longer tail on the low end of the distribution.
2. Three of the average scores differed substantially from the others. These turn out to be 218 (Virgin Islands), 229 (District of Columbia), and 230 (Guam). These three scores could be described as outliers. It is interesting to note that the three unusual values are from the areas that are not states.
3. There do not appear to be any outliers on the high side.
4. A "typical" average math score for the 37 states would be somewhere around 260.
5. There is quite a bit of variability in average score from state to state.

Copyright 2025 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s). Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

Figure 3.55
Stem-and-leaf display and histogram for math test scores.

21H	8
22L	
22H	9
23L	0
23H	
24L	
24H	79
25L	014
25H	6667779999
26L	0003344
26H	55778
27L	12233
27H	667
28L	01

Stem: Tens
Leaf: Ones



How would the displays have been different if the two territories and the District of Columbia had not participated in the testing? The resulting histogram is shown in Figure 3.56. Notice that the display is now more symmetric, with no noticeable outliers. The display still reveals quite a bit of state-to-state variability in average score, and 260 still looks reasonable as a “typical” average score.

Now suppose that the two highest values among the 37 states (Montana and North Dakota) had been even higher. The stem-and-leaf display might then look like the one given in Figure 3.57. In this stem-and-leaf display, two values stand out from the main part of the display. This would catch your attention and might cause you to look carefully at these two states to determine what factors may be related to high math scores.

Figure 3.56
Histogram for the modified math score data.

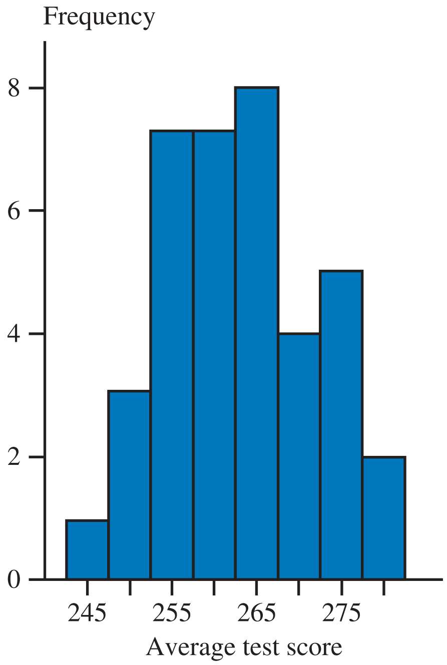


Figure 3.57
Stem-and-leaf display for modified math score data.

24H	79
25L	014
25H	6667779999
26L	0003344
26H	55778
27L	12233
27H	667
28L	
28H	
29L	
29H	68

Stem: Tens
Leaf: Ones

What to Look for in Published Data

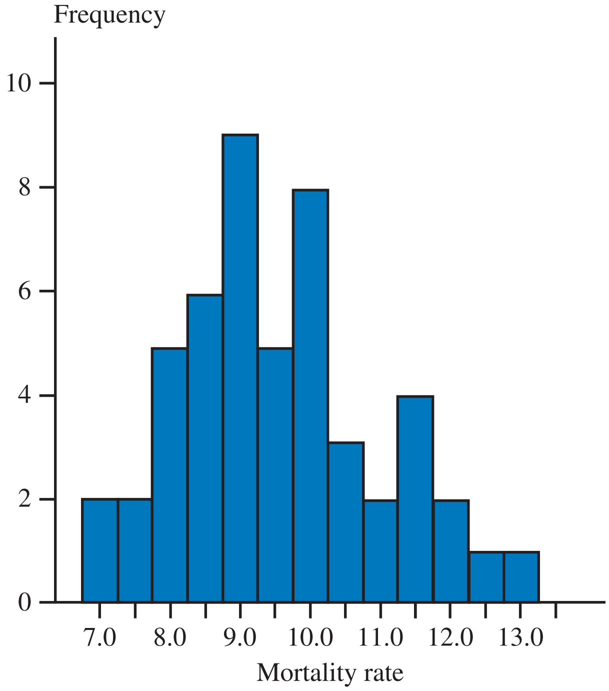
Here are some questions you might ask yourself when attempting to extract information from a graphical data display:

- Is the chosen display appropriate for the type of data collected?
- For graphical displays of univariate numerical data, how would you describe the shape of the distribution, and what does this say about the variable being summarized?
- Are there any outliers (noticeably unusual values) in the data set? Is there any plausible explanation for why these values differ from the rest of the data? (The presence of outliers often leads to further avenues of investigation.)
- Where do most of the data values fall? What is a typical value for the data set? What does this say about the variable being summarized?
- Is there much variability in the data values? What does this say about the variable being summarized?

Of course, you should always think carefully about how the data were collected. If the data were not gathered in a reasonable manner (based on sound sampling methods or experimental design principles), you should be cautious in formulating any conclusions based on the data.

Consider the histogram in Figure 3.58, which is based on data published by the **National Center for Health Statistics**. The data set summarized by this histogram consisted of infant mortality rates (deaths per 1000 live births) for the 50 states in the United States. A histogram is an appropriate way of summarizing these data (although with only 50 observations, a stem-and-leaf display or dotplot would also have been reasonable).

Figure 3.58
Histogram of infant mortality rates.



The histogram itself is slightly positively skewed, with most mortality rates between 7.5 and 12. There is quite a bit of variability in infant mortality rate from state to state—perhaps more than you might have expected. This variability might be explained by differences in economic conditions or in access to health care. You may want to look further into these issues.

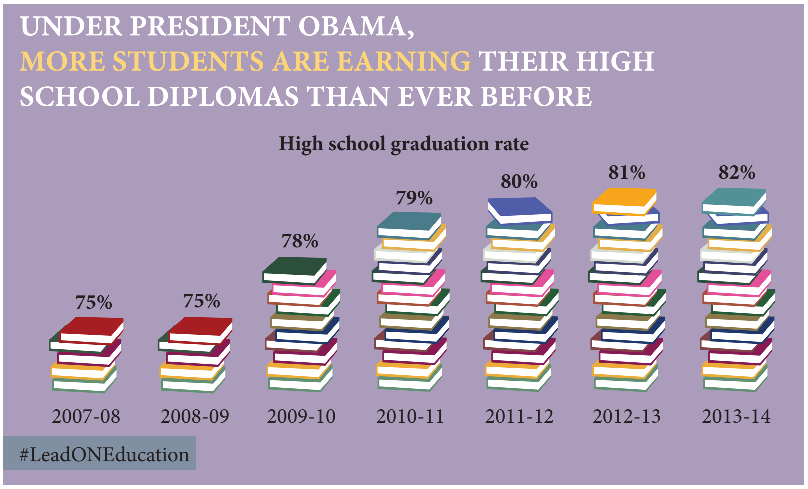
Although there are no obvious outliers, the upper tail is a little longer than the lower tail. The three largest values in the data set are 12.1 (Alabama), 12.3 (Georgia), and 12.8 (South Carolina)—all Southern states. Again, this may suggest some interesting questions that deserve further investigation.

A typical infant mortality rate would be about 9.5 deaths per 1000 live births. This represents an improvement, because researchers at the National Center for Health Statistics stated that the overall rate for 1988 was 10 deaths per 1000 live births. However, they also point out that the United States still ranked 22 out of 24 industrialized nations surveyed, with only New Zealand and Israel having higher infant mortality rates.

A Word to the Wise: Cautions and Limitations

When constructing and interpreting graphical displays, you need to keep in mind these things:

1. **Areas should be proportional to frequency, relative frequency, or magnitude of the number being represented.** The eye is naturally drawn to large areas in graphical displays, and it is natural for the observer to make informal comparisons based on area. Correctly constructed graphical displays, such as pie charts, bar charts, and histograms, are designed so that the areas of the pie slices or the bars are proportional to frequency or relative frequency. Sometimes, in an effort to make graphical displays more interesting, designers lose sight of this important principle, and the resulting graphs are misleading. For example, consider the following graph similar to one that appeared in the online article **"The Most Misleading Charts of 2015, Fixed"** ***(qz.com)***.

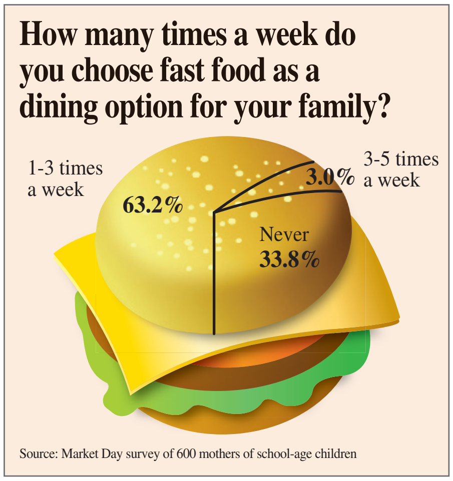


Source: U.S. Department of Education, National Center for Education Statistics.

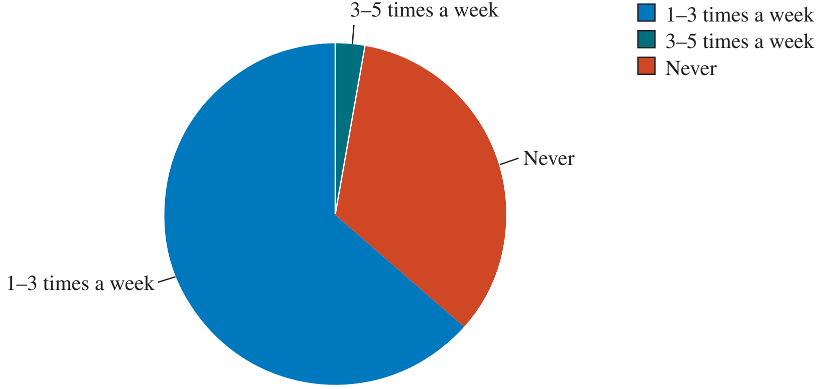
A graph like this appeared in a tweet from the White House in December 2015, and it was accompanied by the statement "Good news: America's high school graduation rate has increased to an all-time high." In trying to make the graph more visually interesting by replacing the bars of a bar chart with stacks of books, areas are distorted and are no longer proportional to the percentages the stacks are representing. Using a five-book sack to represent 75% and a ten-book stack to represent 78% is misleading because 78% is not twice as big as 75%. While the statement that the high school graduation rate had reached an all-time high in the 2013–2014 school year is correct, the graph is misleading in that it exaggerates the size of the change between 2007 and 2014.

Another common distortion occurs when a third dimension is added to bar charts or pie charts. For example, the pie chart at the top left of the next page is similar to one that appeared in ***USA TODAY*** **(September 17, 2009)**. Adding the third dimension distorts the areas and makes it much more difficult to interpret correctly. A correctly drawn pie chart is also shown on the next page.

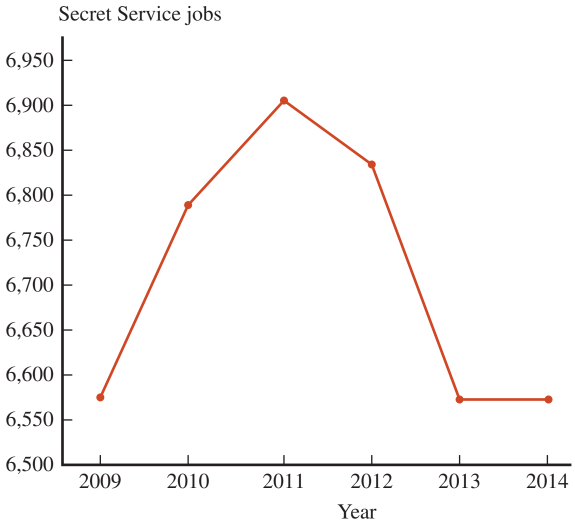
The idea that area should be proportional to frequency, relative frequency, or magnitude of the number that is being represented is discussed in the online article **"The Principle of Proportional Ink" (callingbullshit.org)**. This article is an interesting read and contains a number of revealing examples.



Source: Market Day survey of 600 mothers of school-age children

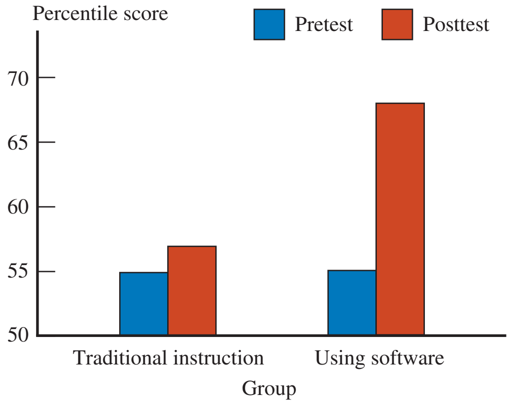


2. **Be cautious of graphs with broken axes.** Although it is common to see scatterplots with broken axes or axes that don't start at 0, be extremely cautious of time series plots, bar charts, or histograms with broken axes or axes that don't start at 0. The use of broken axes in a scatterplot does not distort information about the nature of the relationship in the bivariate data set used to construct the display. On the other hand, in time series plots, broken axes or axes that don't start at 0 can sometimes exaggerate the magnitude of change over time. Although it is not always inadvisable to break the vertical axis in a time series plot, it is something you should watch for, and if you see a time series plot with a broken axis or an axis that doesn't start at 0, as in the accompanying time series plot of the number of people employed by the U.S. Secret Service (similar to a graph that appeared in ***USA TODAY,* October 3, 2014**), you should pay particular attention to the scale on the vertical axis and take extra care in interpreting the graph.



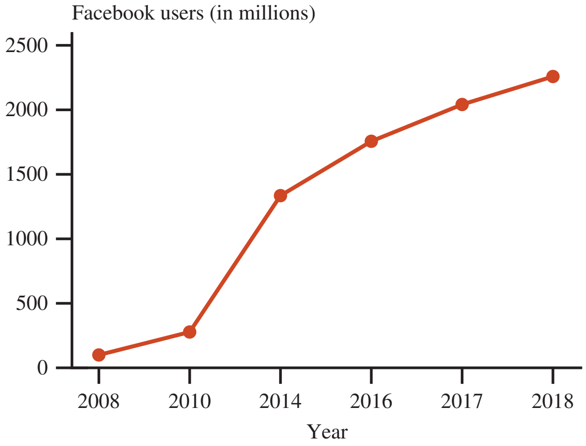
In bar charts and histograms, the vertical axis (which represents frequency, relative frequency, or density) should *never* be broken. If the vertical axis is broken in this type of graph, the resulting display will violate the "proportional area" principle and the display will be misleading. For example, the bar chart at the top of the next page is similar to one appearing in an advertisement for a software product designed to help teachers raise student test scores. By starting the vertical axis at 50, the gain for students using the software is exaggerated. Areas of the bars are not proportional to the magnitude of the numbers represented—the area for the rectangle representing 68 is more than three times the area of the rectangle representing 55!

Copyright 2025 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s). Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

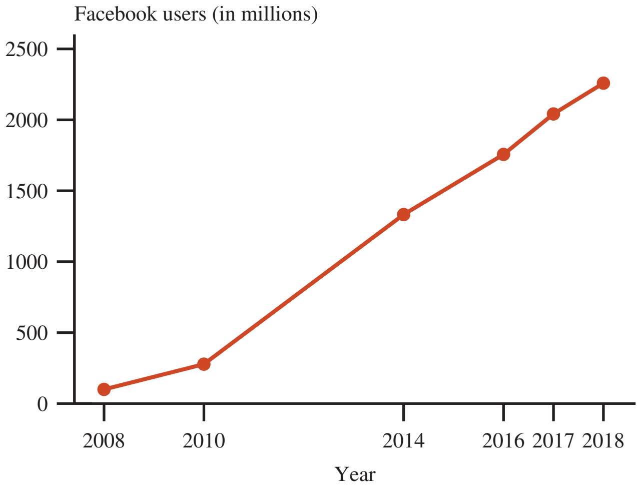


For more information and examples about graphs that are misleading as a result of the axes used, see the online article **"Misleading Axes on Graphs" (callingbullshit.org)**.

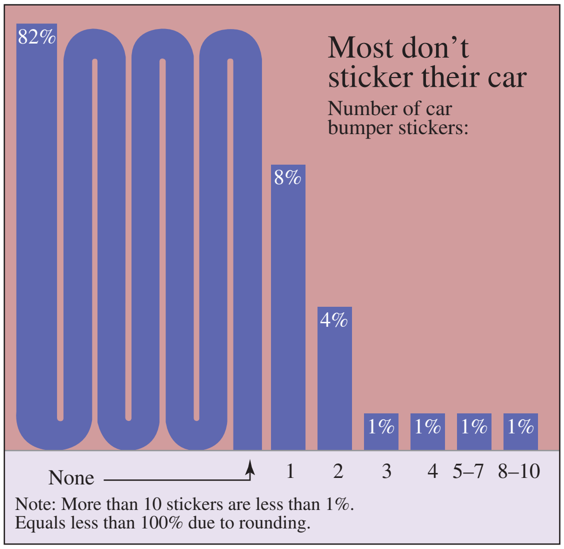
3. **Watch out for unequal time spacing in time series plots.** If observations over time are not made at regular time intervals, special care must be taken in constructing the time series plot. Consider the accompanying time series plot, which is based on data from Our World in Data (ourworldindata.org/rise-of-social-media):



Notice that the intervals between observations are irregular, yet the points in the plot are equally spaced along the time axis. This makes it difficult to make a coherent assessment of the rate of change over time. This could have been remedied by spacing the observations differently along the time axis, as shown in the following plot:



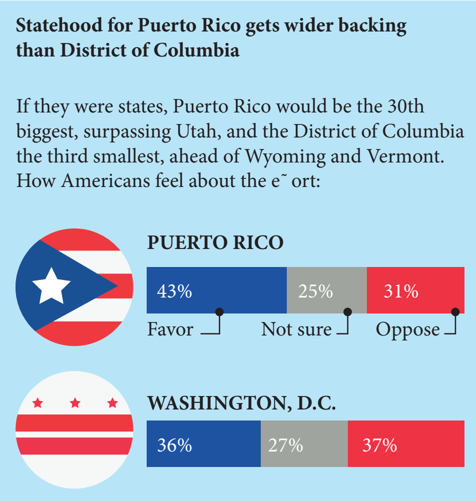
4. **Be careful how you interpret patterns in scatterplots.** A strong pattern in a scatterplot means that the two variables tend to vary together in a predictable way, but it does not mean that there is a cause-and-effect relationship between the two variables. This is considered further in Chapter 5, but in the meantime, when describing patterns in scatterplots, be careful not to use wording that implies that changes in one variable *cause* changes in the other.
5. **Make sure that a graphical display creates the right first impression.** For example, consider the graph on the next page, which is similar to one appearing in ***USA TODAY* (June 25, 2002)**. Although this graph does not violate the proportional area principle, the way the "bar" for the "none" category is displayed makes this graph difficult to read, and a quick glance at this graph would leave the reader with an incorrect impression.



Source: CarMax.com

Exercises 3.61–3.66

3.61 The accompanying graph is similar to one that appeared in ***USA Today* (June 14, 2021)**. It summarizes survey responses to a question about statehood for Puerto Rico and Washington D.C. Suppose that you plan to include this graph in an article that you are writing for your school newspaper. Write a paragraph that could accompany the graph.



Source: U.S. Census Bureau; The Economist/YouGov poll.

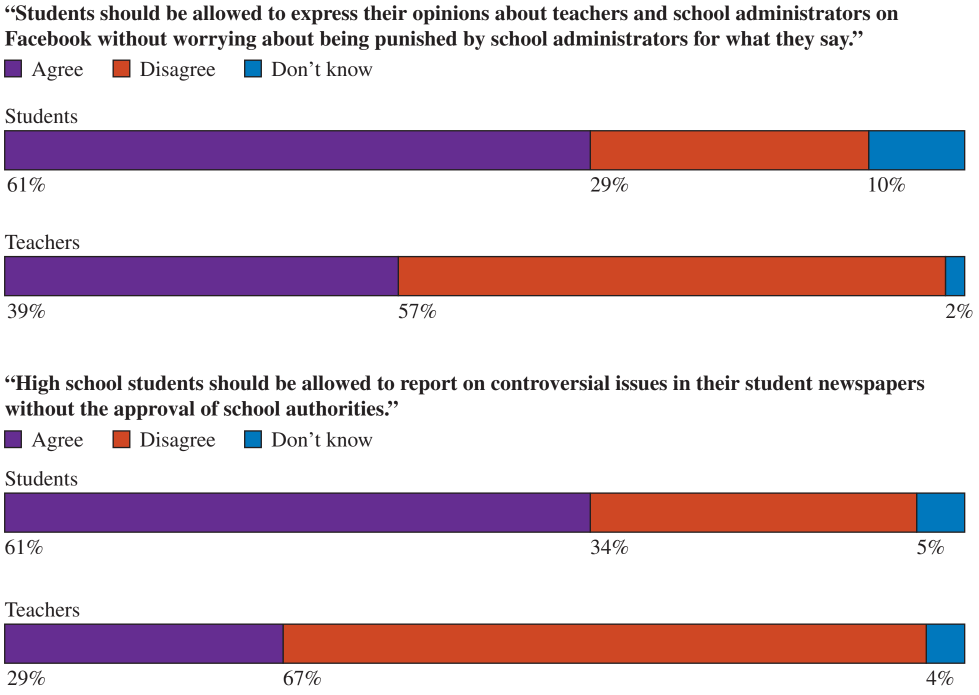
3.62 The segmented bar charts at the top of the next page are similar to ones appearing in the report **"The Future of the First Amendment: 2014 Survey of High School Students and Teachers" (www.knightfoundation.org/)**. The segmented bar charts summarize responses from a large sample of high school students and a large sample of high school teachers.

Suppose that you plan to include these displays in an article that you are writing for your school newspaper. Write a few sentences that could accompany the displays. Be sure to comment on how responses to the two questions are similar or different and on the differences between how teachers and students responded.

3.63 The accompanying bar charts were constructed using data in the article **"Cigarette Smoking Rates Down Sharply Among U.S. Young Adults" (gallup.com, November 28, 2022)**. Each person in large representative samples of adults in the United States (one in 2002 and one in 2022) was asked if they had smoked cigarettes in the past week. The percentages shown in the graph are the percentage who responded yes for each of four different age groups.

Copyright 2025 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s). Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

Figure for Exercise 3.62



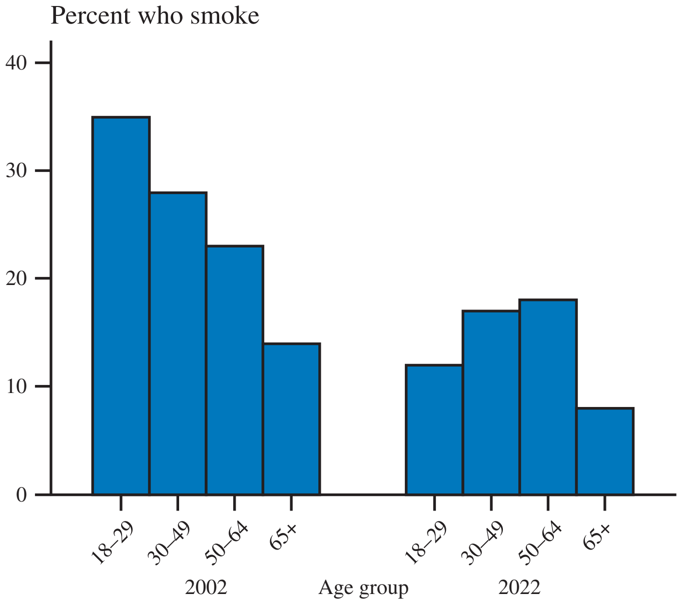
"Students should be allowed to express their opinions about teachers and school administrators on Facebook without worrying about being punished by school administrators for what they say."

"High school students should be allowed to report on controversial issues in their student newspapers without the approval of school authorities."

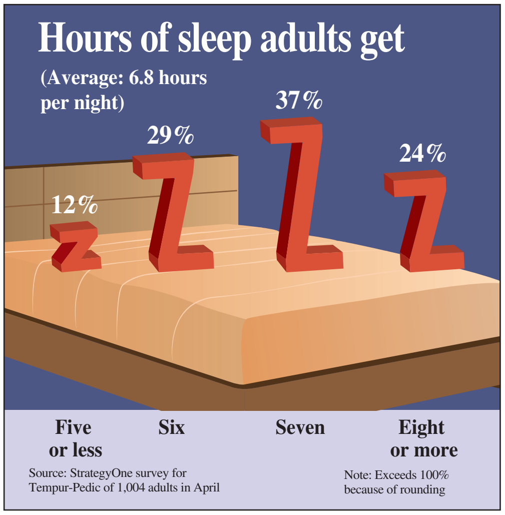
Exercise 3.63 continued

a. Write a few sentences commenting on the change in smoking rates over time.

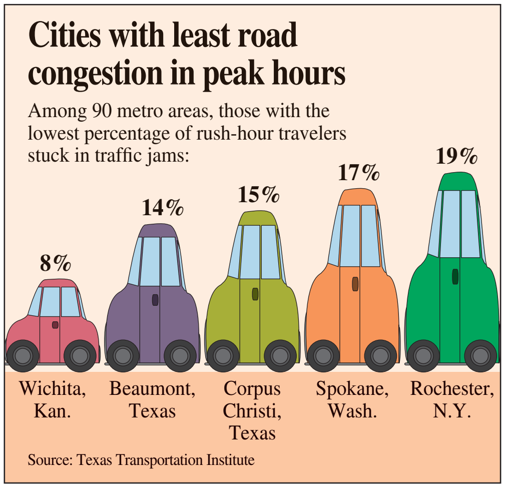
b. Is the conclusion stated in the title of the article reasonable based on the given data? If you could rewrite the title of the article to provide a more complete interpretation of the data, what would you use as a title?



3.64 The two graphical displays to the right are similar to ones that appeared in ***USA TODAY* (June 8, 2009** and **July 28, 2009)**. One is an appropriate representation and the other is not. For each of the two, explain why it is or is not drawn appropriately.



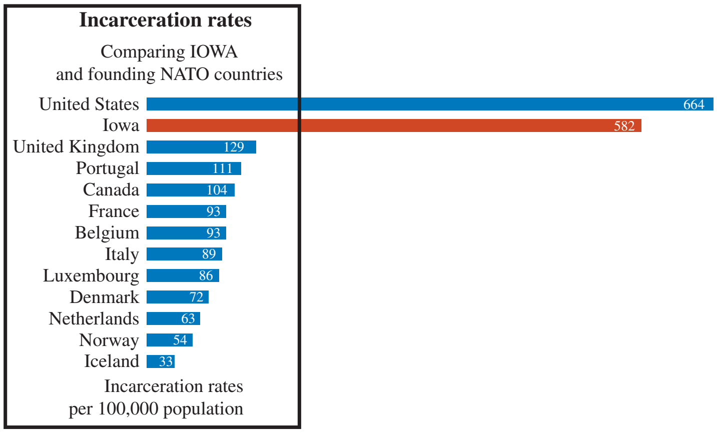
Sources: StrategyOne survey for Tempur-Pedic of 1,004 adults in April; USA TODAY. June 8, 2009.

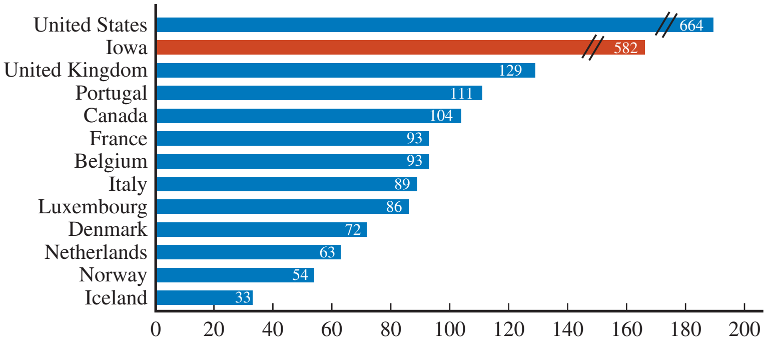


Sources: Texas Transportation Institute; USA TODAY. July 28, 2009.

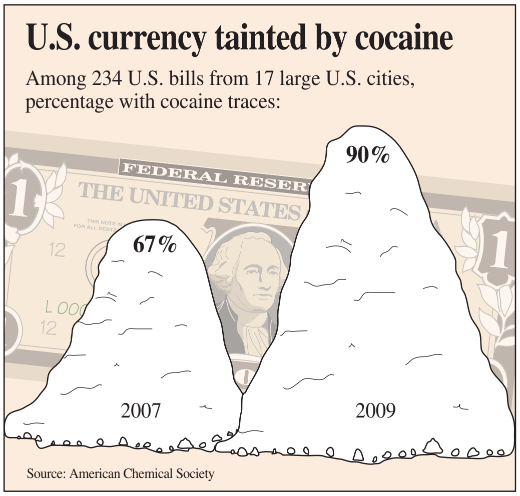
Copyright 2025 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s). Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

3.65 The two graphs below are similar to ones appearing on the policyviz website **(policyviz.com, October 24, 2022)**. Both display data on incarceration rates per 100,000 population. Which of these two graphs is a better representation of the incarceration rates? Explain why the other one is misleading.





3.66 Explain why the following graphical display (similar to one appearing in ***USA TODAY*, September 17, 2009**) is misleading.


U.S. currency tainted by cocaine
Among 234 U.S. bills from 17 large U.S. cities, percentage with cocaine traces:
67%
90%
2007
2009
FEDERAL RESER
THE UNITED STATES

Source: American Chemical Society

Sources: American Chemical Society; USA TODAY. September 17, 2009.

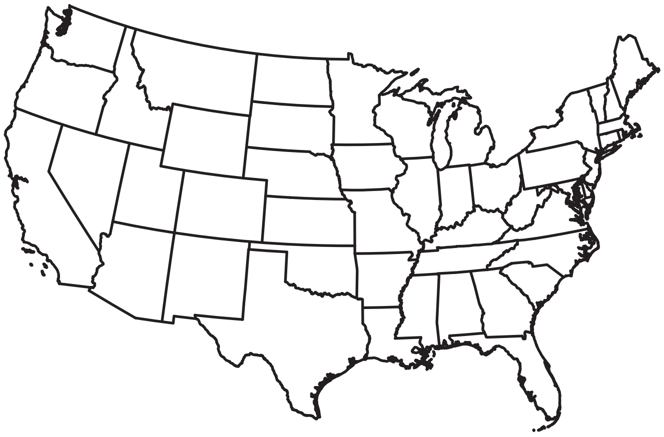
Chapter Activities

Activity 3.1 Locating States

The purpose of this activity is to collect sample data and then summarize the data by constructing and interpreting graphical displays.

Background: A newspaper article bemoaning the state of students' knowledge of geography claimed that more students could identify the island where the 2002 season of the TV show *Survivor* was filmed than could locate Vermont on a map of the United States. In this activity, you will collect data that will allow you to estimate the proportion of students who can correctly locate the states of Vermont and Nebraska.

1. Working as a class, decide how you will select a sample that you think will be representative of the students from your school.
2. Use the sampling method from Step 1 to obtain the subjects for this study. Subjects should be shown the accompanying map of the United States and asked to point out the state of Vermont. After the subject has given their answer, ask the subject to point out the state of Nebraska. For each subject, record whether or not Vermont was correctly identified and whether or not Nebraska was correctly identified.



Copyright 2025 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s). Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

3. When the data collection process is complete, summarize the resulting data in a table like the one shown here:

Response	Frequency
Correctly identified both states	
Correctly identified Vermont but not Nebraska	
Correctly identified Nebraska but not Vermont	
Did not correctly identify either state	

4. Construct a pie chart that summarizes the data in the table from Step 3.
5. What proportion of sampled students were able to correctly identify Vermont on the map?
6. What proportion of sampled students were able to correctly identify Nebraska on the map?
7. Construct a comparative bar chart that shows the proportion correct and the proportion incorrect for each of the two states considered.
8. Which state, Vermont or Nebraska, is closer to the state in which your school is located? Based on the pie chart, do you think that the students at your school were better able to identify the state that was closer than the one that was farther away? Justify your answer.
9. Write a paragraph commenting on the level of knowledge of U.S. geography demonstrated by the students participating in this study.
10. Would you be comfortable generalizing your conclusions in Step 8 to the population of students at your school? Explain why or why not.

Activity 3.2 Bean Counters!

The purpose of this activity is to carry out an experiment and then use graphical displays to summarize the data from the experiment.

Materials needed: A large bowl of dried beans (or marbles, plastic beads, or any other small, fairly regular objects) and a coin.

In this activity, you will investigate whether people can hold more in their right hand or in their left hand.

1. Flip a coin to determine which hand you will measure first. If the coin lands heads side up, start with the right hand. If the coin lands tails side up, start with the left hand. With the designated hand, reach into the bowl and grab as many beans as possible. Raise the hand over the bowl and count to 4. If no beans drop during the count to 4, drop the beans onto a piece of paper and record the number of beans grabbed. If any beans drop during the count, restart the count. That is, you must hold the beans for a count of 4 without any beans falling before you can determine the number grabbed. Repeat the process with the other hand, and then record the following information: (1) right-hand number, (2) left-hand number, and (3) dominant hand (left or right, depending on whether you are left- or right-handed).
2. Create a class data set by recording the values of the three variables listed in Step 1 for each student in your class.
3. Using the class data set, construct a comparative stem-and-leaf display with the right-hand counts displayed on the right and the left-hand counts displayed on the left of the stem-and-leaf display. Comment on the interesting features of the display and include a comparison of the right-hand count and left-hand count distributions.
4. Now construct a comparative stem-and-leaf display that allows you to compare dominant-hand count to nondominant-hand count. Does the display support the theory that dominant-hand count tends to be higher than nondominant-hand count?
5. For each observation in the data set, compute the difference

   dominant-hand count − nondominant-hand count

   Construct a stem-and-leaf display of the differences. Comment on the interesting features of this display.
6. Explain why looking at the distribution of the differences (Step 5) provides more information than the comparative stem-and-leaf display (Step 4). What information is lost in the comparative display that is retained in the display of the differences?

Summary Key Concepts and Formulas

Term or Formula	Comment
Frequency distribution	A table that displays frequencies, and sometimes relative and cumulative relative frequencies, for categories (categorical data), possible values (discrete numerical data), or class intervals (continuous data).
Comparative bar chart	Two or more bar charts that use the same set of horizontal and vertical axes.
Pie chart	A graph of a frequency distribution for a categorical data set. Each category is represented by a slice of the pie, and the area of the slice is proportional to the corresponding frequency or relative frequency.
Segmented bar chart	A graph of a frequency distribution for a categorical data set. Each category is represented by a segment of the bar, and the area of the segment is proportional to the corresponding frequency or relative frequency.
Stem-and-leaf display	A method of organizing numerical data in which the stem values (leading digit(s) of the observations) are listed in a column, and the leaf (trailing digit(s)) for each observation is then listed beside the corresponding stem. Sometimes stems are repeated to stretch the display.
Histogram	A graph of the information in a frequency distribution for a numerical data set. A rectangle is drawn above each possible value (discrete data) or class interval. The rectangle's area is proportional to the corresponding frequency or relative frequency.
Histogram shapes	A (smoothed) histogram may be unimodal (a single peak), bimodal (two peaks), or multimodal. A unimodal histogram may be symmetric, positively skewed (a long right or upper tail), or negatively skewed. A frequently occurring shape is one that is approximately normal.
Cumulative relative frequency plot	A graph of a cumulative relative frequency distribution.
Scatterplot	A graph of bivariate numerical data in which each observation (x, y) is represented as a point with respect to a horizontal x-axis and a vertical y-axis.
Time series plot	A graphical display of numerical data collected over time.

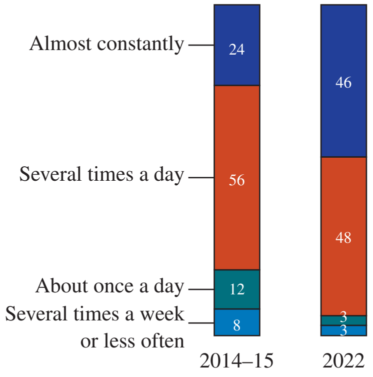
Chapter Review Exercises 3.67–3.73

● Data set available online

3.67 The accompanying segmented bar charts are similar to ones appearing in the report **"Teens, Social Media and Technology 2022"** (Pew Research Center, August 2022). The segmented bar charts summarize responses from large representative samples of teens age 13 to 17 in surveys conducted in 2014–2015 and 2022.

Nearly half of teens now say they use the internet 'almost constantly'

% of U.S. teens who say they use the internet ...



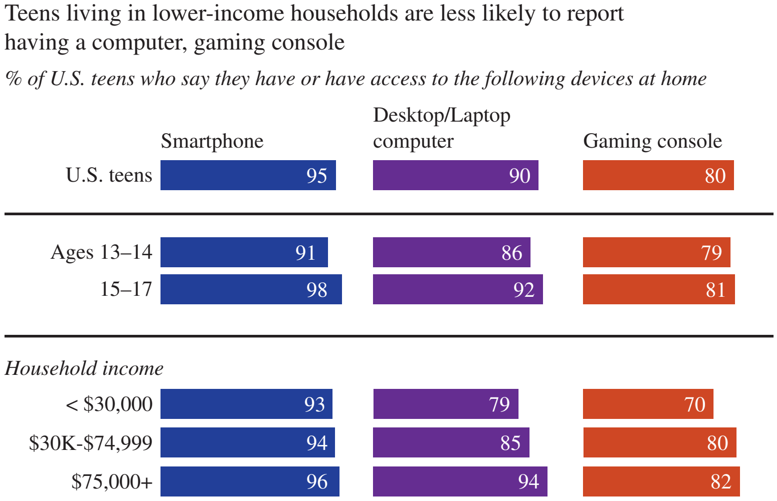
Suppose that you plan to include these segmented bar graphs in an article that you are writing for your school newspaper. Write a few sentences that could accompany the displays. Be sure to comment on any notable differences between 2014–15 and 2022.

3.68 The report referenced in the previous exercise also included a graph similar to the one shown at the top of the next page. How is the title of the graph supported by what you see in the data summarized in the graph?

3.69 ● The data in the table on the next page are the percentage of drivers who are uninsured in each of the 50 U.S. states and the District of Columbia in 2020 as reported in the article **"Uninsured Motorists Statistics" (valuepenguin.com)**. Construct a graphical display that shows the distribution of percent uninsured drivers and write a few sentences commenting on what the display reveals about the distribution.

Copyright 2025 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s). Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

Figure for Exercise 3.68



Teens living in lower-income households are less likely to report having a computer, gaming console

% of U.S. teens who say they have or have access to the following devices at home

Data for exercise 3.69

State	Percent Uninsured Drivers	State	Percent Uninsured Drivers
Alabama	19.5	Missouri	16.4
Alaska	16.1	Montana	8.5
Arizona	11.8	Nebraska	9.3
Arkansas	19.3	Nevada	10.4
California	16.6	New Hampshire	6.1
Colorado	16.3	New Jersey	3.1
Connecticut	6.3	New Mexico	21.8
Delaware	8.5	New York	4.1
District of Columbia	19.1	North Carolina	7.4
Florida	20.4	North Dakota	13.0
Georgia	12.4	Ohio	13.0
Hawaii	9.3	Oklahoma	13.4
Idaho	13.2	Oregon	10.7
Illinois	11.8	Pennsylvania	6.0
Indiana	15.8	Rhode Island	16.5
Iowa	11.3	South Carolina	10.9
Kansas	10.9	South Dakota	7.4
Kentucky	13.9	Tennessee	23.7
Louisiana	11.7	Texas	8.3
Maine	4.9	Utah	6.5
Maryland	14.1	Vermont	8.5
Massachusetts	3.5	Virginia	10.5
Michigan	25.5	Washington	21.7
Minnesota	9.9	West Virginia	9.2
Mississippi	29.4	Wisconsin	13.3
		Wyoming	5.8

3.70 The report **"The American Freshman: National Norms Fall 2019" (Higher Education Research Institute, 2019)** summarizes data collected from more than 95,000 college freshmen across the United States. One question on the survey asked students how they identified politically, with choices of far left, liberal, middle-of-the-road, conservative, and far right. The data are summarized in the accompanying table.

Response	Relative Frequency
Far left	0.045
Liberal	0.322
Middle-of-the-road	0.436
Conservative	0.178
Far right	0.019

a. Use the given data to construct a bar chart. Write a brief description of how the surveyed students identified politically that is based on the bar chart.

b. Another question on the survey asked students how likely they were to vote in a future local, state, or national election. The responses to the voting question are summarized for each of the five political identities in the accompanying table. Construct a comparative bar chart and compare the voting responses for the five political identification choices.

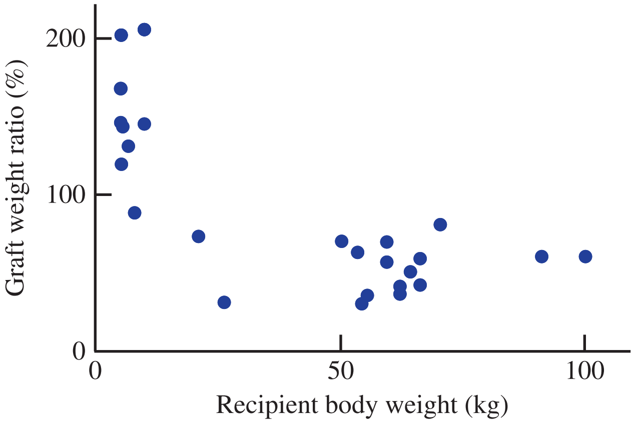
	Relative Frequency				
	Political Identification				
	Far Left	Liberal	Middle-of-the-Road	Conservative	Far Right
Likely to Vote	0.80	0.76	0.53	0.68	0.65
Unlikely to Vote	0.20	0.24	0.47	0.32	0.35

3.71 Does the size of a transplanted organ matter? A study that attempted to answer this question **("Minimum Graft Size for Successful Living Donor Liver Transplantation,"** ***Transplantation*** **[1999]: 1112–1116)** included a graph similar to the scatterplot at the top of the next page. "Graft weight ratio" is the weight of the transplanted liver relative to the ideal size liver for the recipient.

Copyright 2025 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s). Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

a. Discuss interesting features of this scatterplot.

b. Why do you think the overall relationship is negative?



3.72 ● Each year Gallup surveys people from around the world about their positive and negative daily experiences. The survey data are then used to calculate a global Positive Experience Index. Possible values for this index are between 0 and 100, with higher scores indicating greater overall positive experiences. The values in the accompanying table are from the report **“Gallup Global Emotions 2020” (www.gallup.com)**.

Year	Positive Experience Index
2006	68
2007	68
2008	69
2009	70
2010	70
2011	69
2012	69
2013	71
2014	71
2015	71
2016	70
2017	69
2018	71
2019	71

(continued)

a. Construct a time series plot that shows how the positive experience index has changed over time.

b. Has the positive experience index increased at a steady rate? Write a few sentences commenting on what you see in the time series plot.

3.73 ● The Education Data Initiative (educationdata.org) reported the following data on average cost per year for tuition and fees for four-year public universities in the United States. Construct a time series plot of these data and comment on the trend over time.

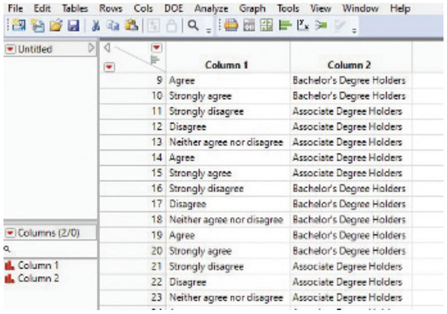
Year	Average Cost
2006	$5,351
2007	$5,666
2008	$5,943
2009	$6,312
2010	$6,717
2011	$7,132
2012	$7,713
2013	$8,070
2014	$8,312
2015	$8,543
2016	$8,778
2017	$8,804
2018	$9,036
2019	$9,212
2020	$9,349

Technology Notes

Comparative Bar Charts

JMP

1. Enter the raw data into one column
2. In a separate column enter the group information



	Column 1	Column 2
9	Agree	Bachelor's Degree Holders
10	Strongly agree	Bachelor's Degree Holders
11	Strongly disagree	Associate Degree Holders
12	Disagree	Associate Degree Holders
13	Neither agree nor disagree	Associate Degree Holders
14	Agree	Associate Degree Holders
15	Strongly agree	Associate Degree Holders
16	Strongly disagree	Bachelor's Degree Holders
17	Disagree	Bachelor's Degree Holders
18	Neither agree nor disagree	Bachelor's Degree Holders
19	Agree	Bachelor's Degree Holders
20	Strongly agree	Bachelor's Degree Holders
21	Strongly disagree	Associate Degree Holders
22	Disagree	Associate Degree Holders
23	Neither agree nor disagree	Associate Degree Holders

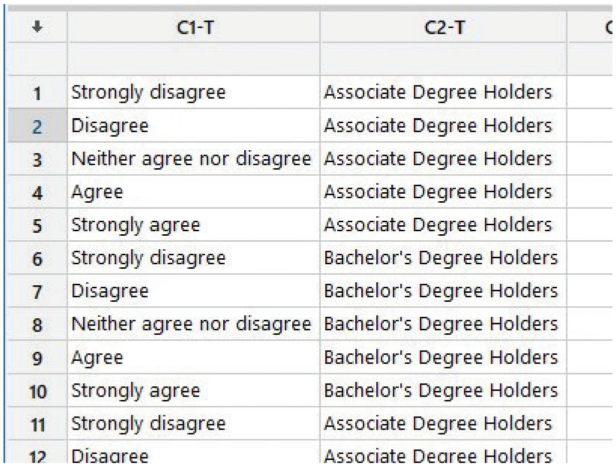
Source: JMP

3. Click **Graph, Legacy** and then select **Chart**
4. Click and drag the column containing the raw data from the box under **Select Columns** to the box next to **Categories, X, Levels**
5. Click and drag the column containing the group information from the box under **Select Columns** to the box next to **Categories, X, Labels**
6. Click **OK**

Minitab

1. Input the group information into C1
2. Input the raw data into C2 (be sure to match the response from the first column with the appropriate group)

Copyright 2025 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s). Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.



	C1-T	C2-T
1	Strongly disagree	Associate Degree Holders
2	Disagree	Associate Degree Holders
3	Neither agree nor disagree	Associate Degree Holders
4	Agree	Associate Degree Holders
5	Strongly agree	Associate Degree Holders
6	Strongly disagree	Bachelor's Degree Holders
7	Disagree	Bachelor's Degree Holders
8	Neither agree nor disagree	Bachelor's Degree Holders
9	Agree	Bachelor's Degree Holders
10	Strongly agree	Bachelor's Degree Holders
11	Strongly disagree	Associate Degree Holders
12	Disagree	Associate Degree Holders

Source: Minitab

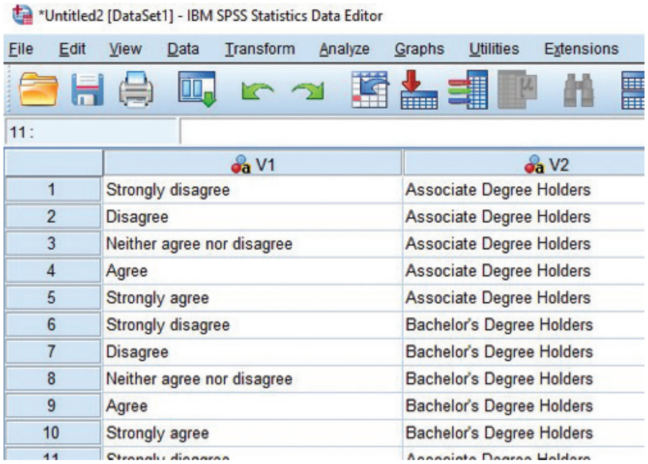
3. Select **Graph** then choose **Bar Chart…**
4. Highlight Cluster
5. Click **OK**
6. Double click C1 to add it to the **Categorical Variables** box
7. Then double click C2 to add it to the **Categorical Variables** box
8. Click **OK**

Note: Be sure to list the grouping variable first in the Categorical Variables box.

Note: You may add or format titles, axis titles, legends, and so on, by clicking on the **Labels…** button prior to performing Step 8 above.

SPSS

1. Enter the raw data into one column
2. Enter the group information into a second column (be sure to match the response from the first column with the appropriate group)


*Untitled2 [DataSet1] - IBM SPSS Statistics Data Editor
File
Edit
View
Data
Transform
Analyze
Graphs
Utilities
Extensions
11 :

	V1	V2
1	Strongly disagree	Associate Degree Holders
2	Disagree	Associate Degree Holders
3	Neither agree nor disagree	Associate Degree Holders
4	Agree	Associate Degree Holders
5	Strongly agree	Associate Degree Holders
6	Strongly disagree	Bachelor's Degree Holders
7	Disagree	Bachelor's Degree Holders
8	Neither agree nor disagree	Bachelor's Degree Holders
9	Agree	Bachelor's Degree Holders
10	Strongly agree	Bachelor's Degree Holders
11	Strongly disagree	Associate Degree Holders

Source: SPSS

3. Select **Graph** and choose **Chart Builder…**
4. Under **Choose from** highlight Bar
5. Click and drag the second bar chart in the first column (Clustered Bar) to the Chart preview area
6. Click and drag the group variable (second column) into the **Cluster on X** box in the upper right corner of the chart preview area
7. Click and drag the data variable (first column) into the **X-Axis?** box in the chart preview area

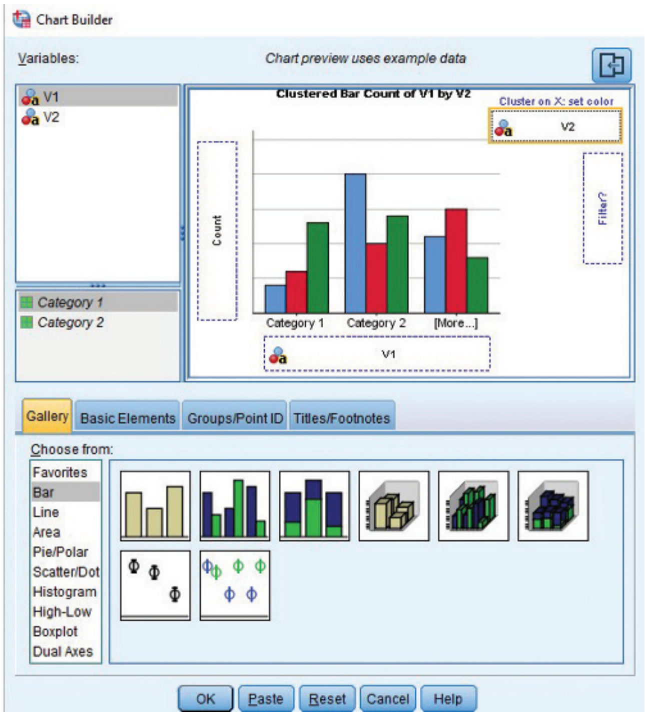


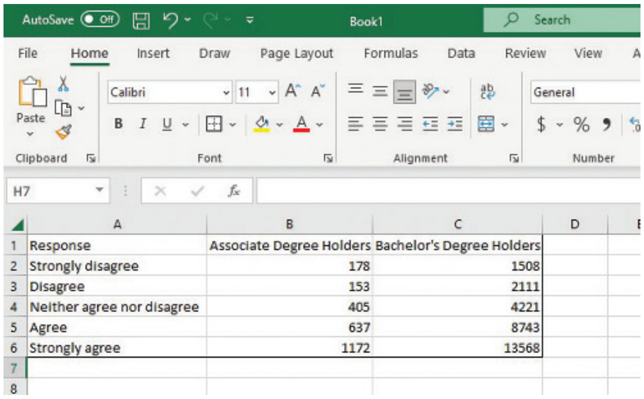
Chart Builder
Variables:
Chart preview uses example data
V1
V2
Clustered Bar Count of V1 by V2
Cluster on X: set color
V2
Count
Filter?
Category 1
Category 2
[More...]
V1
Category 1
Category 2
Gallery
Basic Elements
Groups/Point ID
Titles/Footnotes
Choose from:
Favorites
Bar
Line
Area
Pie/Polar
Scatter/Dot
Histogram
High-Low
Boxplot
Dual Axes
OK
Paste
Reset
Cancel
Help

Source: SPSS

8. Click **Ok**

Excel

1. Enter the category names into column A (you may input the title for the variable in cell A1)
2. Enter the count or percent for different groups into a new column, starting with column B
3. Select all data (including column titles if used)


AutoSave
Off
Book1
Search
File
Home
Insert
Draw
Page Layout
Formulas
Data
Review
View
Paste
Calibri
11
General
Clipboard
Font
Alignment
Number
H7
A
B
C
D

	A	B	C
1	Response	Associate Degree Holders	Bachelor's Degree Holders
2	Strongly disagree	178	1508
3	Disagree	153	2111
4	Neither agree nor disagree	405	4221
5	Agree	637	8743
6	Strongly agree	1172	13568
7			
8			

Source: Microsoft Corporation

4. Click on the **Insert** Ribbon
5. Choose **Column** and select the first chart under 2-D Column (Clustered Column)
6. The chart will appear on the same worksheet as your data

Note: You may add or format titles, axis titles, legends, and so on, by right-clicking on the appropriate piece of the chart.

Note: Using the Bar Option on the Insert Ribbon will produce a horizontal bar chart.

Copyright 2025 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s). Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

TI-83/84

The TI-83/84 does not have the functionality to produce comparative bar charts.

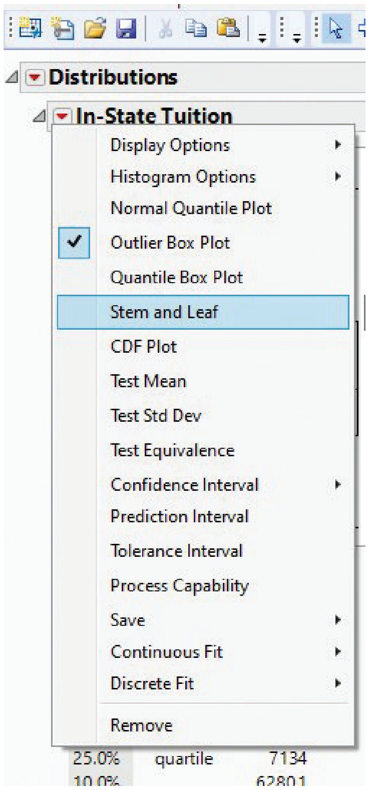
TI-Nspire

The TI-Nspire does not have the functionality to produce comparative bar charts.

Stem-and-Leaf Plots

JMP

1. Enter the raw data into a column
2. Click **Analyze** then select **Distribution**
3. Click and drag the column name containing the data from the box under **Select Columns** to the box next to **Y, Columns**
4. Click **OK**
5. Click the red arrow next to the column name


Distributions
In-State Tuition
Display Options
Histogram Options
Normal Quantile Plot
Outlier Box Plot
Quantile Box Plot
Stem and Leaf
CDF Plot
Test Mean
Test Std Dev
Test Equivalence
Confidence Interval
Prediction Interval
Tolerance Interval
Process Capability
Save
Continuous Fit
Discrete Fit
Remove
25.0% quartile 7134
10.0% 62801

Source: JMP

6. Select **Stem and Leaf**

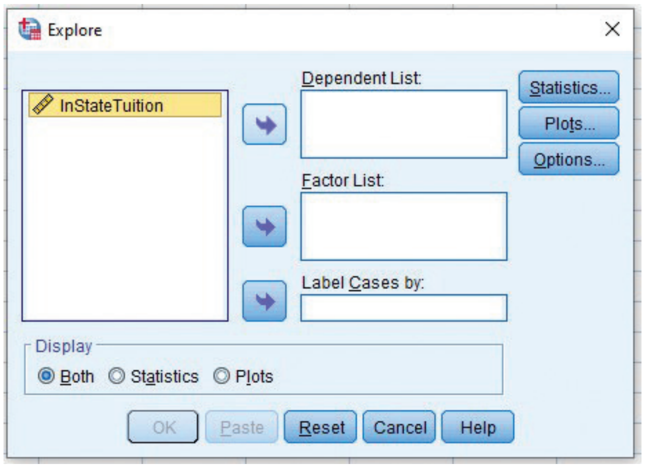
Minitab

1. Input the raw data into C1
2. Select **Graph** and choose **Stem-and-leaf…**
3. Double click C1 to add it to the **Graph Variables** box
4. Click **OK**

Note: You may add or format titles, axis titles, legends, and so on, by clicking on the **Labels…** button prior to performing Step 4 above.

SPSS

1. Input the raw data into a column
2. Select **Analyze** and choose **Descriptive Statistics** then **Explore**
3. Highlight the variable name from the box on the left
4. Click the arrow to the right of the Dependent List box to add the variable to this box


Explore
InStateTuition
Dependent List:
Factor List:
Label Cases by:
Statistics...
Plots...
Options...
Display
● Both ○ Statistics ○ Plots
OK
Paste
Reset
Cancel
Help

Source: SPSS

5. Click **OK**

Note: The stem-and-leaf plot is one of the plots produced along with several other descriptive statistics and plots.

Excel

Excel does not have the functionality to create stem-and-leaf plots automatically.

TI-83/84

The TI-83/84 does not have the functionality to create stem-and-leaf plots.

TI-Nspire

The TI-Nspire does not have the functionality to create stem-and-leaf plots.

Histograms

JMP

1. Enter the raw data into a column
2. Click **Analyze** then select **Distribution**
3. Click and drag the column name containing the data from the box under **Select Columns** to the box next to **Y, Columns**
4. Click **OK**
5. Click the red arrow next to the column name
6. Select **Histogram Options** and click **Vertical**

Minitab

1. Input the raw data into C1
2. Select **Graph** and choose **Histogram…**
3. Highlight Simple
4. Click **OK**
5. Double click C1 to add it to the **Graph Variables** box
6. Click **OK**

Note: You may add or format titles, axis titles, legends, and so on, by clicking on the **Labels…** button prior to performing Step 5 above.

SPSS

1. Input the raw data into a column
2. Select **Graph** and choose **Chart Builder…**

Copyright 2025 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s). Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

3. Under **Choose from** highlight Histogram
4. Click and drag the first option (Simple Histogram) to the Chart preview area
5. Click and drag the variable name from the **Variables** box into the **X-Axis?** box
6. Click **OK**

Excel

Note: In order to produce histograms in Excel, we will use the Data Analysis Add-On. For instructions on installing this add-on, please see the note at the end of this chapter's Technology Notes.

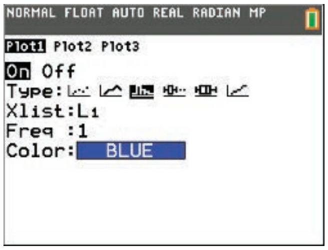
1. Input the raw data in column A (you may use a column heading in cell A1)
2. Click on the **Data** ribbon
3. Choose the **Data Analysis** option from the **Analysis** group
4. Select **Histogram** from the Data Analysis dialog box and click **OK**
5. Click in the **Input Range:** box and then click and drag to select your data (including the column heading)
    a. If you used a column heading, click the checkbox next to **Labels**
6. Click the checkbox next to **Chart Output**
7. Click **OK**

Note: If you have specified bins manually in another column, click in the box next to **Bin Range:** and select your bin assignments.

Note: You may adjust the space between the bars by right-clicking on one of the bars and making Gap Width 0%.

TI-83/84

1. Enter the data into **L1** (In order to access lists press the **STAT** key, highlight the option called **Edit…** then press **ENTER**)
2. Press the **2nd** key then the **Y =** key
3. Highlight the option for Plot1 and press **ENTER**
4. Highlight **On** and press **ENTER**
5. For **Type**, select the histogram shape and press **ENTER**



On Off
Type:
Xlist:L1
Freq :1
Color: BLUE

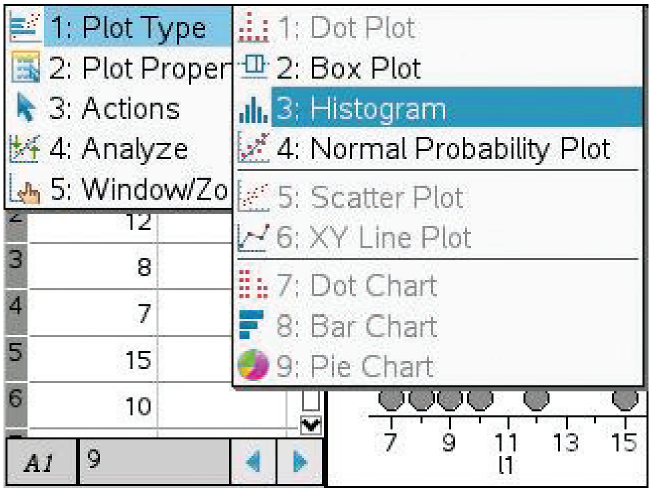
6. Press **ZOOM** then select **9:ZoomStat** to automatically resize the window and display the graph.

TI-Nspire

1. Enter the data into a data list (In order to access data lists select the spreadsheet option and press **enter**)

Note: Be sure to title the list by selecting the top row of the column and typing a title.

2. Press the **menu** key then select **3:Data** then select **9:Quick Graph** and press **enter** (a dotplot will appear)
3. Press the **menu** key and select **1:Plot Type**
4. Select **3:Histogram** and press **enter**



Source: TI-Nspire

Scatterplots

JMP

1. Input the data for the independent variable into the first column
2. Input the data for the dependent variable into the second column
3. Click **Analyze** then select **Fit Y by X**
4. Click and drag the column name for the in dependent data from the box under **Select Columns** to the box next to **Y, Response**
5. Click and drag the column name for the independent data from the box under **Select Columns** to the box next to **X, Factor**
6. Click **OK**

Minitab

1. Input the raw data for the independent variable into C1
2. Input the raw data for the dependent variable into C2
3. Select **Graph** and choose **Scatterplot**
4. Highlight Simple
5. Click **OK**
6. Double click on C2 to add it to the **Y Variables** column of the spreadsheet
7. Double click on C1 to add it to the **X Variables** column of the spreadsheet
8. Click **OK**

Note: You may add or format titles, axis titles, legends, and so on, by clicking on the **Labels…** button prior to performing Step 7 above.

SPSS

1. Input the raw data into two columns
2. Select **Graph** and choose **Chart Builder…**
3. Under **Choose from** highlight Scatter/Dot
4. Click and drag the first option (Simple Scatter) to the Chart preview area
5. Click and drag the variable name representing the independent variable to the **X-Axis?** box
6. Click and drag the variable name representing the dependent variable to the **Y-Axis?** box
7. Click **OK**

Copyright 2025 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s). Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

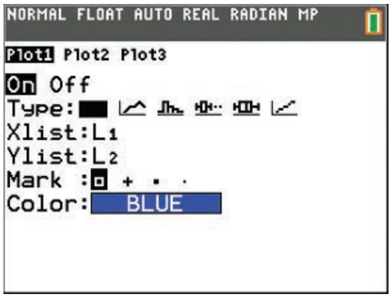
Excel

1. Input the raw data into two columns (you may enter column headings)
2. Select both columns of data
3. Click on the **Insert** ribbon
4. Click **Scatter** and select the first option (Scatter with Markers Only)

Note: You may add or format titles, axis titles, legends, and so on, by right-clicking on the appropriate piece of the chart.

TI-83/84

1. Input the data for the dependent variable into L2 (In order to access lists press the **STAT** key, highlight the option called **Edit…** then press **ENTER**)
2. Input the data for the independent variable into L1
3. Press the **2nd** key then press the **Y =** key
4. Select the **Plot1** option and press **ENTER**
5. Select **On** and press **ENTER**
6. Select the scatterplot option and press **ENTER**



On Off
Type:
Xlist:L1
Ylist:L2
Mark :
Color: BLUE

7. Press **ZOOM** then select **9:ZoomStat** to automatically resize the window and display the graph.

TI-Nspire

1. Enter the data for the independent variable into a data list (In order to access data lists select the spreadsheet option and press **enter**)

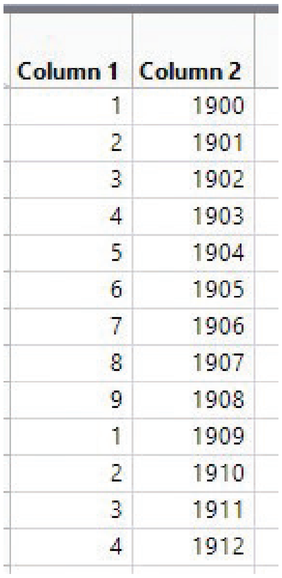
Note: Be sure to title the list by selecting the top row of the column and typing a title.

2. Enter the data for the dependent variable into a separate data list
3. Highlight both columns of data (by scrolling to the top of the screen to highlight one list then press shift and use the arrow key to highlight the second list)
4. Press the **menu** key and select **3:Data** then select **9:Quick Graph** and press **enter**

Time Series Plots

JMP

1. Input the raw data into one column
2. Input the time increment data into a second column
3. Click **Analyze** then **Quality and Process (updated)** then select **Legacy Control Charts** followed by **Runs Chart**
4. Click and drag the name of the column containing the raw data from the box under **Select Columns** to the box next to **Process**



Column 1	Column 2
1	1900
2	1901
3	1902
4	1903
5	1904
6	1905
7	1906
8	1907
9	1908
1	1909
2	1910
3	1911
4	1912

Source: JMP

5. Click and drag the name of the column containing the time increment data from the box under **Select Columns** to the box next to **Sample Label**
6. Click **OK**

Minitab

1. Input the raw data into C1
2. Select **Graph** and choose **Time Series Plot…**
3. Highlight Simple
4. Click **OK**
5. Double click C1 to add it to the **Series** box
6. Click the Time/Scale button
7. Choose the appropriate time scale under Time Scale (Choose Calendar if you want to use Days, Years, Months, Quarters, etc.)
8. Click to select one set for each variable
9. In the spreadsheet, fill in the start value for the time scale
10. Input the increment value into the box next to Data box next to **Data Increment**: (i.e. **Increment:** (i.e., 1 to move by one year or one month, etc.)
11. Click **OK**
12. Click **OK**

Note: You may add or format titles, axis titles, legends, and so on, by clicking on the **Labels…** button prior to performing Step 12 above.

SPSS

1. Input the raw data into two columns: one representing the data and one representing the time increments
2. Select **Analyze** then choose **Forecasting** then **Sequence Charts…**
3. Highlight the column name containing the raw data and add it to the **Variables** box
4. Highlight the column name containing the time data and add it to the **Time Axis Labels** box
5. Click **OK**

Excel

1. Input the data into two columns: one representing the data and one representing the time increments
2. Select an empty cell

3. Click on the **Insert** ribbon
4. Click **Insert Line or Area Chart** and select the first option under 2-D Line (Line)
5. Right-click on the empty chart area that appears and select **Select Data…**
6. Under **Legend Entries (Series)** click **Add**
7. In the **Series name:** box select the column title for the data
8. In the **Series values:** box select the data values
9. Click **OK**
10. Under **Horizontal (Category) Axis Labels** click **Edit**
11. In the **Axis label range:** select the time increment data (do NOT select the column title).
12. Click **OK**
13. Click **OK**

TI-83/84

1. Input the time increment data into **L1**
2. Input the data values for each time into **L2**
3. Press the **2nd** key then the **Y =** key
4. Select **1:Plot1** and press **ENTER**
5. Highlight **On** and press **ENTER**
6. Highlight the second option in the first row and press **ENTER**
7. Press **ZOOM** then select **9:ZoomStat** to automatically resize the window and display the graph.

TI-Nspire

1. Enter the data for the time increments into a data list (In order to access data lists select the spreadsheet option and press **enter**)

Note: Be sure to title the list by selecting the top row of the column and typing a title.

2. Enter the data for each time increment into a separate data list
3. Highlight both columns of data (by scrolling to the top of the screen to highlight one list then press shift and use the arrow key to highlight the second list)
4. Press the **menu** key and select **3:Data** then select **9: Quick Graph** and press **enter**
5. Press the **menu** key and select **1:Plot Type** then select **6:XY Line Plot**

Installing Excel's Data Analysis Add-On

1. Click **File** and then Click **Options**
2. Click **Add-Ins**, then in the **Manage** box, select **Excel Add-ins**
3. Click **Go**
4. In the **Add-Ins available** box, select the **Analysis ToolPak** checkbox, and then click **OK**

Note: If this option is not listed, click **Browse** to locate it.

Note: If you get prompted that the Analysis ToolPak is not currently installed on your computer, click **Yes** to install it.

5. After you load the Analysis ToolPak, the **Data Analysis** command is available in the **Analysis** group on the **Data** ribbon.

Cumulative Review Exercises CR3.1–CR3.16

● Data set available online

CR3.1 For each of the following, determine whether the statistical study described is an observational study or an experiment. Give a brief explanation of your choice.

a. Can choosing the right music make wine taste better? This question was investigated by a researcher at a university in Edinburgh **(www.decanter.com)**. Each of 250 volunteers was assigned at random to one of five rooms where they were asked to taste and rate a glass of wine. No music was playing in one of the rooms, and a different style of music was playing in each of the other four rooms. The researcher concluded that cabernet sauvignon is rated more highly when bold music is played than when no music is played.

b. The article **"Sugar-Sweetened Beverage Intake in Adulthood and Adolescence and Risk of Early Onset Colorectal Cancer among Women" (Gut.bmj.com, 2021)** analyzed data from 95,464 female registered nurses who participated in a prospective health study over a 25-year period. As part of the study, the nurses reported their consumption of sugar-sweetened beverages using food frequency questionnaires every four years. The study found that nurses who consumed more than two servings of sugar-sweetened beverages had more than twice the risk of colon cancer compared to those who consumed less than one sugar-sweetened beverage per week.

c. ***USA TODAY* (January 29, 2003)** reported that in a study of affluent Americans (defined as those with incomes of $75,000 or more per year) 57% indicated that they would rather have more time than more money.

d. The article **"Acupuncture for Bad Backs: Even Sham Therapy Works" (*Time*, May 12, 2009)** summarized a study conducted by researchers at the Group Health Center for Health Studies in Seattle. In this study, 638 adults with back pain were randomly assigned to one of four groups. People in group 1 received the usual care for back pain. People in group 2 received acupuncture at a set of points tailored specifically for each individual. People in group 3 received acupuncture at a standard set of points typically used in the treatment of back pain. Those in group 4 received fake acupuncture—they were poked with a toothpick at the same set of points used for the people in group 3. Two notable conclusions from the study were: (1) patients receiving real or fake acupuncture experienced a greater reduction in pain than those receiving usual care; and (2) there was no significant difference in pain reduction between those who received real acupuncture (groups 2 and 3) and those who received fake acupuncture toothpick pokes.

CR3.2 The article **"Sugar-Sweetened Beverage Intake in Adulthood and Adolescence and Risk of Early Onset Colorectal Cancer among Women" (Gut.bmj.com, 2021)** analyzed data from 95,464 female registered nurses who participated in a prospective health study over a 25-year period. As part of the study, the nurses reported their consumption of sugar-sweetened beverages using food frequency questionnaires every four years. Explain how measurement (response) bias could occur in a study that uses food frequency questionnaires.

CR3.3 Sometimes samples are composed entirely of volunteer responders. Give a brief description of the dangers of using voluntary response samples.

CR3.4 In a ***Freakonomics*** podcast episode titled **"Should America be Run by Trader Joe's?" (November 28, 2018)**, a researcher described an experiment conducted at a Northern California grocery store to investigate how giving customers lots of choices would affect their behavior. Specifically, "she set up a tasting booth for jams. And she alternated the choice set: sometimes the booth would feature six different jams and sometimes 24." She found that "a larger choice set generates more interest [stopping at the booth], but the smaller choice set generates more action [purchasing jam]." Do you think that alternating the size of the choice set (6 jams or 24 jams) is close enough to random assignment? That is, do you think it would tend to create comparable experimental groups?

CR3.5 An article called **"The Placebo Effect Works and You Can Catch It From Your Doctor" (*NPR*, 2019)** describes a study to investigate whether patients can pick up subtle facial cues that reveal their doctor's beliefs about the effectiveness of a treatment. The article included this study description:

> Researchers randomly assigned undergraduate students to play the role of a patient or a doctor. The "patients" were given a controlled heat stimulus to the forearm, after receiving one of two types of cream from the "doctor." Students in the doctor group had previously been conditioned to believe that one of the creams was pain reliever. But in reality both of the two creams that they administered were an identical petroleum-jelly based placebo. And yet, when the doctor actors believed that the cream was a real medication … the patient actors reported experiencing significantly lower amounts of pain.

a. The article also includes the statement, "Avoiding these placebo effects is one reason for double blind studies." Explain what the term *double-blind* means and discuss the purpose of blinding in an experiment.

b. Given how these data were collected, do you think it is reasonable to conclude that the doctors' belief about the effectiveness of the pain reliever is what *caused* the observed differences in patient reported pain?

c. Given how these data were collected, do you think it is reasonable to generalize the results of this study to all doctors and patients?

CR3.6 To examine the relationship between noise exposure and heart disease, a study **("Exposure to Loud Noise, Bilateral High-Frequency Hearing Loss and Coronary Heart Disease," *Occupational and Environmental Medicine*, 2016)** compared a group of people with hearing loss to a group of people without hearing loss. The researchers found that the percentage of people in the hearing loss group who had heart disease was higher than the percentage for the group that did not have hearing loss. This association was particularly strong for workers exposed to loud occupational noise.

a. Is this study an experiment? If so, what is the explanatory variable, and what is the response variable? If not, explain why it is not an experiment.

b. From this study alone, is it reasonable to conclude that noise exposure is the cause of the observed difference in heart disease risk? Justify your answer.

CR3.7 In a survey of 100 people who had recently purchased motorcycles, data on the following variables were recorded:

Gender of purchaser
Brand of motorcycle purchased
Number of previous motorcycles owned by purchaser
Telephone area code of purchaser
Weight of motorcycle as equipped at purchase

a. Which of these variables are categorical?

b. Which of these variables are discrete numerical?

c. Which type of graphical display would be an appropriate choice for displaying the distribution of gender, a bar chart or a dotplot?

d. Which type of graphical display would be an appropriate choice for displaying the distribution of weight, a bar chart or a dotplot?

CR3.8 **EducationData.org** reported the percentage of students who had defaulted on a student loan within 1 year of when they were scheduled to begin repayment **(Student Loan Default Rate, educationdata.org, July 10, 2021)**. Information was given for public colleges, private nonprofit colleges, and private for-profit colleges.

	Relative Frequency		
Loan Status	**Public Colleges**	**Private Nonprofit Colleges**	**Private For-Profit Colleges**
Good Standing	0.892	0.929	0.844
In Default	0.108	0.071	0.156

a. Construct a comparative bar chart that would allow you to compare loan status for the three types of colleges.

b. The website states "Student borrowers who attended private for-profit colleges are the most likely to default on

Copyright 2025 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s). Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

their loans, whereas private non-profit college attendees are the least likely to default." What aspect of the comparative bar chart supports this statement?

CR3.9 ● The article **"Fliers Trapped on Tarmac Push for Rules on Release" (*USA TODAY*, July 28, 2009)** gave the following data for 17 airlines on the number of flights that were delayed on the tarmac for at least 3 hours for the period between October 2008 and May 2009:

Airline	Number of Delays	Rate per 10,000 Flights
ExpressJet	93	4.9
Continental	72	4.1
Delta	81	2.8
Comair	29	2.7
American Eagle	44	1.6
US Airways	46	1.6
JetBlue	18	1.4
American	48	1.3
Northwest	24	1.2
Mesa	17	1.1
United	29	1.1
Frontier	5	0.9
SkyWest	29	0.8
Pinnacle	13	0.7
Atlantic Southeast	11	0.6
AirTran	7	0.4
Southwest	11	0.1

The figure at the top of the next page shows two dotplots: one displays the number of delays data, and one displays the rate per 10,000 flights data.

a. If you were going to rank airlines based on flights delayed on the tarmac for at least 3 hours, would you use the *total number of flights* data or the *rate per 10,000 flights data*? Explain the reason for your choice.

b. Write a short paragraph that could be part of a newspaper article on flight delays that could accompany the dotplot of the *rate per 10,000 flights* data.

CR3.10 ● The Sharecare Community Well-Being Index combines information on more than 600 health risk factors (**wellbeingindex.sharecare.com/**). The 2020 index values for each of the 50 U.S. states are given in the accompanying table. Higher scores indicate greater well-being

a. Construct a stem-and-leaf display of these data.

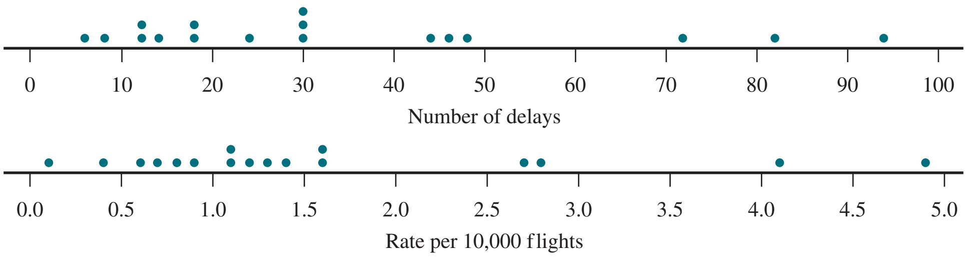
b. Based on the stem-and-leaf display, what do you notice about the center and variability of the data distribution?

c. Do any values in the data set stand out as unusual? If so, which states correspond to the unusual observations and how do these values differ from the rest?

State	Community Well-Being Index
Alabama	52
Alaska	59
Arizona	58
Arkansas	49
California	67
Colorado	65
Connecticut	66
Delaware	60
Florida	59
Georgia	56
Hawaii	71
Idaho	55
Illinois	63
Indiana	54
Iowa	56
Kansas	58
Kentucky	50
Louisiana	52
Maine	57
Maryland	70
Massachusetts	71
Michigan	55
Minnesota	62
Mississippi	46
Missouri	56
Montana	56
Nebraska	60
Nevada	59
New Hampshire	61
New Jersey	70
New Mexico	49
New York	69
North Carolina	56
North Dakota	60
Ohio	56
Oklahoma	51
Oregon	63
Pennsylvania	61
Rhode Island	64
South Carolina	55
South Dakota	57
Tennessee	53
Texas	56
Utah	65
Vermont	59
Virginia	64
Washington	65
West Virginia	50
Wisconsin	58
Wyoming	53

Copyright 2025 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s). Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

Figure for Exercise CR 3.9



CR3.11 The report **"Trends in College Pricing 2015" (trends.collegeboard.org/)** included the information in the accompanying relative frequency distributions for public and private nonprofit four-year college students.

- **a.** Construct a relative frequency histogram for tuition and fees for students at public four-year colleges. Write a few sentences describing the distribution of tuition and fees, commenting on shape, center, and variability.
- **b.** Construct a relative frequency histogram for tuition and fees for students at private nonprofit four-year colleges. Be sure to use the same scale for the vertical and horizontal axes as you used for the histogram in Part (a). Write a few sentences describing the distribution of tuition and fees for students at private nonprofit four-year colleges.
- **c.** Write a few sentences describing the differences in the two distributions.

	Proportion of Students (Relative Frequency)	
Tuition and Fees	**Public Four-Year College Students**	**Private Nonprofit Four-Year College Students**
\$3,000 to < \$6,000	0.04	0.02
\$6,000 to < \$9,000	0.38	0.06
\$9,000 to < \$12,000	0.27	0.02
\$12,000 to < \$15,000	0.12	0.02
\$15,000 to < \$18,000	0.04	0.03
\$18,000 to < \$21,000	0.04	0.03
\$21,000 to < \$24,000	0.02	0.05
\$24,000 to < \$27,000	0.02	0.07
\$27,000 to < \$30,000	0.02	0.09
\$30,000 to < \$33,000	0.02	0.09
\$33,000 to < \$36,000	0.01	0.08
\$36,000 to < \$39,000	0.00	0.09
\$39,000 to < \$42,000	0.00	0.07
\$42,000 to < \$45,000	0.00	0.07
\$45,000 to < \$48,000	0.00	0.21

CR3.12 ● The paper **"Lessons from Pacemaker Implantations" (*Journal of the American Medical Association* [1965]: 231–232)** gave the results of a study that followed 89 heart patients who had received electronic pacemakers. The time (in months) to the first electrical malfunction of the pacemaker was recorded:

24	20	16	32	14	22	2	12	24	6	10	20
8	16	12	24	14	20	18	14	16	18	20	22
24	26	28	18	14	10	12	24	6	12	18	16
34	18	20	22	24	26	18	2	18	12	12	8
24	10	14	16	22	24	22	20	24	28	20	22
26	20	6	14	16	18	24	18	16	6	16	10
14	18	24	22	28	24	30	34	26	24	22	28
30	22	24	22	32							

- **a.** Summarize these data in the form of a frequency distribution, using class intervals of 0 to < 6, 6 to < 12, and so on.
- **b.** Calculate the relative frequencies and cumulative relative frequencies for each class interval of the frequency distribution of Part (a).
- **c.** Show how the relative frequency for the class interval 12 to < 18 could be obtained from the cumulative relative frequencies.
- **d.** Use the cumulative relative frequencies to give approximate answers to the following:
  - i. What proportion of those who participated in the study had pacemakers that did not malfunction within the first year?
  - ii. If the pacemaker must be replaced as soon as the first electrical malfunction occurs, approximately what proportion required replacement between 1 and 2 years after implantation?
- **e.** Construct a cumulative relative frequency plot, and use it to answer the following questions.
  - i. What is the approximate time at which 50% of the pacemakers had failed?
  - ii. What is the approximate time at which only 10% of the pacemakers initially implanted were still functioning?

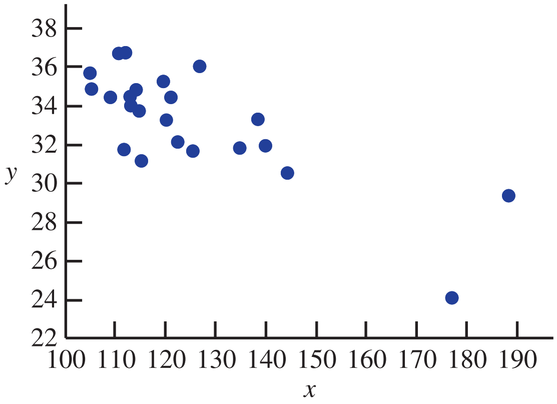
CR3.13 ● **The Nilson Report (nilsonreport.com, December 2020)** included the data in the accompanying table on the amount of money lost to credit card fraud for the years from 2013 to 2019 and projected losses for 2021 to 2027.

Copyright 2025 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s). Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

Construct a time series plot for these data and describe the trend over time. Is the amount of money lost to credit card fraud increasing at a fairly steady rate? Explain.

Year	Money Lost to Credit Card Fraud (in billions of dollars)
2013	13.7
2015	21.8
2017	24.0
2019	28.7
2021	32.0
2023	33.0
2025	35.3
2027	38.5

CR3.14 One factor in the development of tennis elbow, a malady that strikes fear into the hearts of all serious players of that sport, is the impact-induced vibration of the racket-and-arm system at ball contact. It is well known that the likelihood of getting tennis elbow depends on various properties of the racket used. Consider the accompanying scatterplot of x = racket resonance frequency (in hertz) and y = sum of peak-to-peak accelerations (a characteristic of arm vibration, in meters per second per second) for $n = 23$ different rackets **("Transfer of Tennis Racket Vibrations into the Human Forearm," *Medicine and Science in Sports and Exercise* [1992]: 1134–1140)**. Discuss interesting features of the data and of the scatterplot.



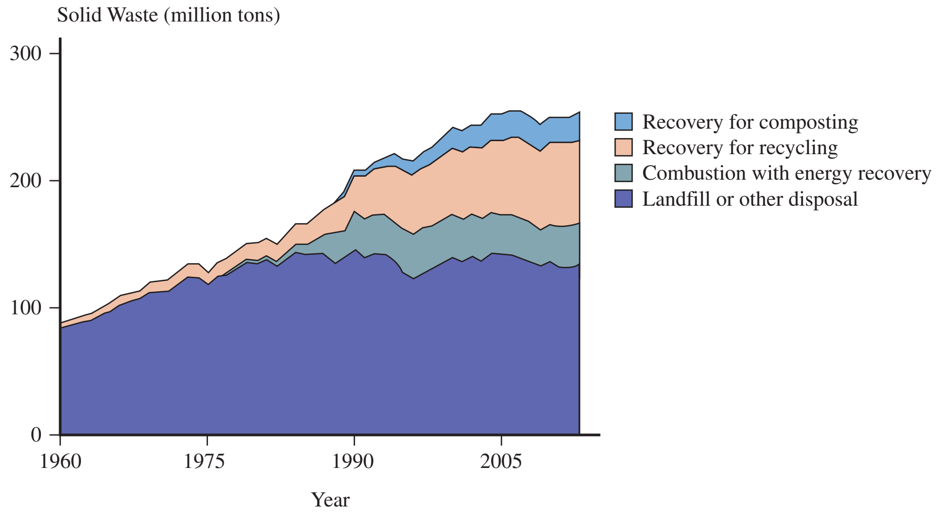
CR3.15 The Solid Waste Management section of the Environmental Protection Agency ***Report on the Environment* (www.epa.gov/roe/)** included a graph similar to the graph at the bottom of the page. The report also included the following statement:

> The last several decades have seen steady growth in recycling and composting, while the total amounts landfilled peaked in 1990 (145 MT) and have generally declined since then (134 MT in 2013).

Explain how the time series plot is consistent or is not consistent with the given statement.

Figure for CR 3.15

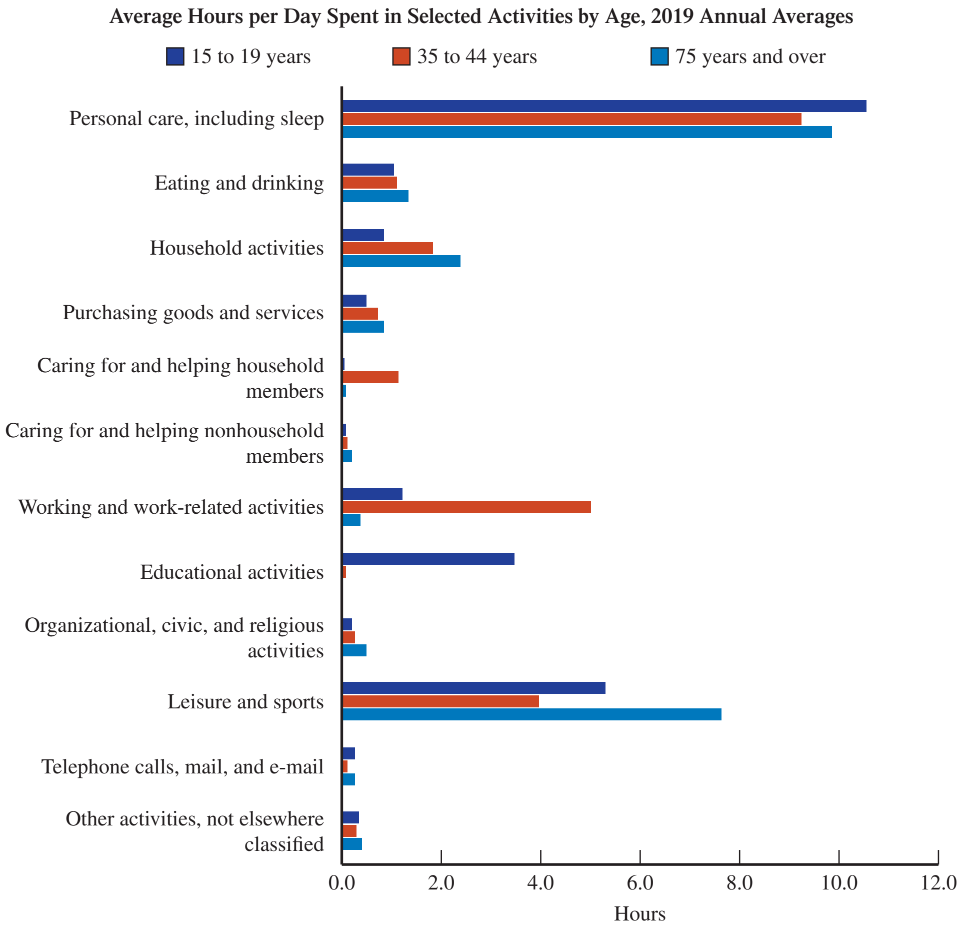
Exhibit 1. Municipal solid waste generated and managed in the U.S., 1960–2013



CR3.16 The accompanying graphical display was generated using the interactive tool that appeared on the website of the U.S. Bureau of Labor Statistics **(www.bls.gov)**. To make this graph, three age groups were selected for the display.

a. What do you notice about the relationships between the variables in the graph?

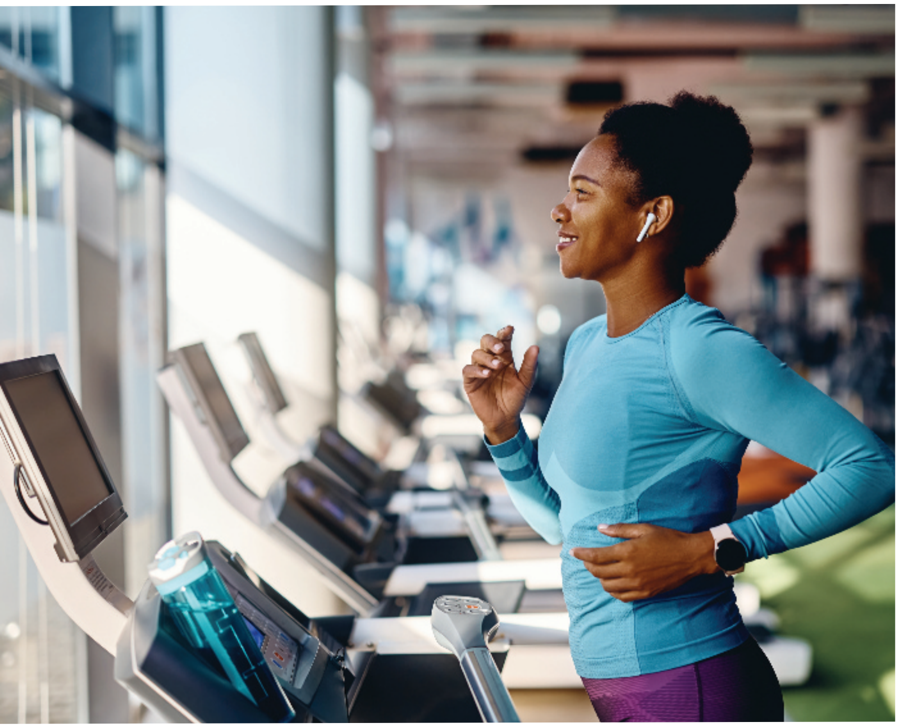
b. What story does the graphical display tell?



Copyright 2025 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s). Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

Chapter 4

Numerical Methods for Describing Data



Drazen Zigic/Shutterstock.com

Preview Example: Just Thinking About Exercise?

Studies have shown that in some cases people overestimate the extent to which physical exercise can compensate for food consumption. When this happens, people increase food intake more than what is justified based on the exercise performed. The authors of the paper **"Just Thinking About Exercise Makes Me Serve More Food: Physical Activity and Calorie Compensation" (*Appetite* [2011]: 332–335)** wondered if even *thinking* about exercise would lead to increased food consumption.

They carried out an experiment in which people were offered snacks as a reward for participating in the experiment. People read a short essay and then answered a few questions about the essay. Some participants read an essay that was unrelated to exercise (the control group), some read an essay that described listening to music while taking a 30-minute walk (the fun group), and some read an essay that described strenuous exercise (the exercise group).

Participants were then given two plastic bags and invited to help themselves to two types of snacks—Chex Mix and M&Ms. After the participants served themselves, the bags were weighed so that the researchers could determine the number of calories in snacks taken.

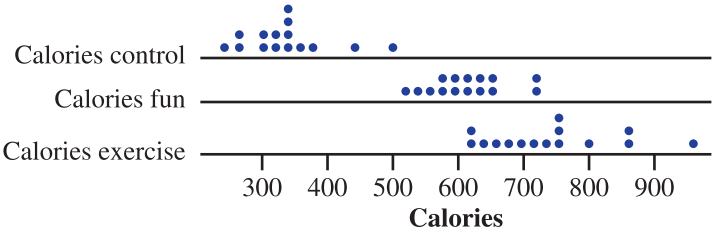
Learning Objectives

After completing this chapter, you will be able to:

- Explain how the variance and standard deviation describe variability in a data set.
- Describe the impact that outliers can have on measures of center and spread.
- Explain the difference between a sample statistic and a population characteristic.
- Calculate and interpret the values of the sample mean and the sample median.
- Calculate and interpret the values of the sample standard deviation and the interquartile range.
- Construct and interpret a boxplot.
- Identify outliers in numerical data.
- Make statements about a data distribution using Chebyshev's Rule and the Empirical Rule.
- Describe relative standing using percentiles and z scores.

Data on number of calories consistent with summary values in the paper were used to construct the comparative dotplot shown in Figure 4.1. From the dotplots, it is clear that the number of calories consumed differs from person to person and also tends to be quite a bit higher for those who read about exercise than for those in the control group! To further compare the three distributions, it is helpful to summarize them numerically.

Figure 4.1
Comparative dotplot of calories.



In this chapter, you will learn how to calculate numerical summary measures that describe both the center and the extent of variability in a data set.

Section 4.1 Describing the Center of a Data Set

When describing numerical data, it is common to report a value that is representative of the observations in the data set. Such a number describes roughly where the data are located or "centered" along the number line, and is called a measure of center. The two most widely used measures of center are the mean and the median.

The Mean

The **mean** of a numerical data set is just the familiar arithmetic average: the sum of the observations divided by the number of observations. At this point, it is helpful to introduce notation for the variable on which observations are made, for the number of observations in the data set, and for the individual observations:

x = the variable observed
n = the number of observations in the data set (the sample size)
x_1 = the first observation in the data set
x_2 = the second observation in the data set
⋮
x_n = the nth (last) observation in the data set

For example, you might have a sample consisting of $n = 4$ observations on x = time to complete an online hotel reservation (in minutes):

$$x_1 = 5.9 \qquad x_2 = 7.3 \qquad x_3 = 6.6 \qquad x_4 = 5.7$$

Notice that the value of the subscript on x has no relationship to the magnitude of the observation. In this example, x_1 is just the first observation in the data set and not necessarily the smallest observation, and x_n is the last observation but not necessarily the largest.

The sum of $x_1, x_2, \ldots, x_n$ can be denoted by $x_1 + x_2 + \cdots + x_n$, but this can also be written using summation notation. The Greek letter Σ denotes summation. In particular, Σx denotes the sum of all the x values in the data set under consideration.*

---

*It is also common to see Σx written as Σx_i or even as $\sum_{i=1}^{n} x_i$, but for simplicity we will usually omit the summation indices.

Copyright 2025 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s). Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

Definition

Sample mean: The **sample mean** of a sample consisting of numerical observations $x_1, x_2, \ldots, x_n$ is denoted by $\bar{x}$, and its formula is given by

$$\bar{x} = \frac{\text{sum of all observations in the sample}}{\text{number of observations in the sample}} = \frac{x_1 + x_2 + \cdots + x_n}{n} = \frac{\Sigma x}{n}$$

Example 4.1 Thinking About Exercise Again

Understand the context ➤

● The example in the chapter introduction described a study that investigated how thinking about exercise might affect food intake. Data on calories in snacks taken by people in each of three groups (control, read about fun walk, and read about strenuous exercise) are given in Table 4.1. Dotplots of these three data sets were given in the chapter introduction and are reproduced here as Figure 4.2.

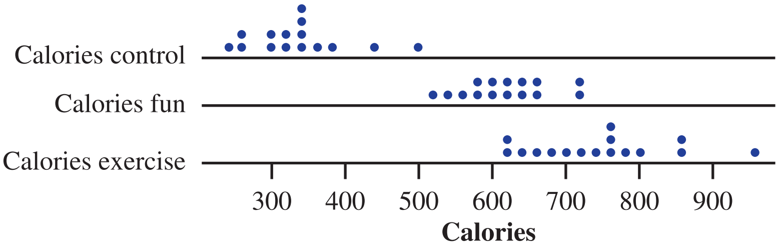


Figure 4.2
Dotplot of calories.

What Do You Think?

Which of the three groups has the greatest sample mean? Which has the smallest sample mean?

Table 4.1 Calorie Data

Consider the data ➤

Control Group	Fun Group	Exercise Group
340	668	626
300	585	802
329	637	768
381	588	738
331	529	751
445	597	715
320	719	670
256	612	630
252	622	862
342	553	761
332	600	648
296	658	956
357	542	671
505	711	854
242	641	703

Do the work ➤

For the control group, the sum of the sample data values is

$$\Sigma x = 340 + 300 + \cdots + 242 = 5028$$

and the sample mean number of calories is

$$\bar{x} = \frac{\Sigma x}{n} = \frac{5028}{15} = 335.2$$

● **Data set available online**

Interpret the results ➤ The value of the sample mean describes where the number of calories data for the control group is centered along the number line. It can be interpreted as a typical number of calories for people in the control group. For the fun group, the sample mean number of calories is

$$\bar{x} = \frac{\Sigma x}{n} = \frac{9262}{15} = 617.5$$

and for the exercise group the sample mean number of calories is

$$\bar{x} = \frac{\Sigma x}{n} = \frac{11{,}155}{15} = 743.7$$

Notice that the sample mean number of calories for the control group is much smaller than the means for the other two groups.

The data values in Example 4.1 were all integers, yet the sample means were given as 335.2, 617.5, and 743.7. It is common to use more digits of decimal accuracy for the sample mean. This allows the value of the mean to fall between possible observable values (for example, the average number of children per family could be 1.8, whereas no single family will have 1.8 children). When calculating the value of the sample mean, it is common to round the calculated value to one more decimal place than the data. For example, if the data are integers, the value of the sample mean would be rounded to one decimal place. If the observations in the data set were numbers with one decimal place, the value of the sample mean would be rounded to two decimal places.

The sample mean $\bar{x}$ is calculated using sample data, so it is a characteristic of that particular sample. It is customary to use Roman letters to denote sample characteristics, as we have done with $\bar{x}$. Characteristics of the population are usually denoted by Greek letters. One of the most important population characteristics is the population mean.

Definition

Population mean: The **population mean**, denoted by μ, is the average of all x values in the entire population.

For example, the average fuel efficiency for *all* 600,000 cars of a particular model might be $\mu = 27.52$ mpg. A sample of $n = 5$ cars might yield efficiencies of 27.3, 26.2, 28.4, 27.9, 26.5, with $\bar{x} = 27.26$ for this particular sample (somewhat smaller than μ). However, a second sample might result in $\bar{x} = 28.52$, a third $\bar{x} = 26.85$, and so on. The value of $\bar{x}$ varies from sample to sample, whereas there is just one value for μ.

In later chapters, you will see how the value of $\bar{x}$ from a particular sample can be used to draw various conclusions about the value of μ. Example 4.2 illustrates how the value of $\bar{x}$ from a particular sample can differ from the value of μ and how the value of $\bar{x}$ differs from sample to sample.

Example 4.2 County Population Sizes

Understand the context ➤ The 50 states plus the District of Columbia contain 3137 counties. Let x denote the number of residents of a county. Then there are 3137 values of the variable x in the population. The sum of these 3137 values is 231,665,106 (2018 Census Bureau estimate), so the population average value of x is

$$\mu = \frac{231{,}665{,}106}{3137} = 73{,}849.3 \text{ residents per county}$$

The Census Bureau website was used to select three different random samples from this population of counties, with each sample consisting of five counties. The results appear in Table 4.2, along with the sample mean for each sample. The three $\bar{x}$ values are different from one another because they are based on three different samples and the value of $\bar{x}$ depends on the x values in the sample. Also notice that none of the three values is close to the value of the population mean, μ. If you did not know the value of μ, you might use one of these $\bar{x}$ values as an *estimate* of μ, but in each case the estimate would be far from the actual value.

Table 4.2 **Three Samples from the Population of All U.S. Counties (x = number of residents)**

Sample 1		Sample 2		Sample 3	
County	**x Value**	**County**	**x Value**	**County**	**x Value**
Fayette, TX	20,964	Stoddard, MO	28,509	Chattahoochee, GA	21,332
Monroe, IN	101,719	Johnston, OK	10,442	Petroleum, MT	638
Greene, NC	15,584	Sumter, AL	17,002	Armstrong, PA	77,299
Shoshone, ID	19,021	Milwaukee, WI	960,664	Schoolcraft, MI	8,625
Jasper, IN	26,570	Albany, WY	30,607	Benton, MO	12,133
	$\Sigma x = 183{,}858$		$\Sigma x = 1{,}047{,}224$		$\Sigma x = 120{,}027$
	$\bar{x} = 36{,}771.6$		$\bar{x} = 209{,}444.8$		$\bar{x} = 24{,}005.4$

Alternatively, you could combine the three samples into a single sample with $n = 15$ observations:

$$x_1 = 20{,}964, \ldots, x_5 = 26{,}570, \ldots, x_{15} = 12{,}133$$

$$\Sigma x = 1{,}351{,}109$$

$$\bar{x} = \frac{1{,}351{,}109}{15} = 90{,}073.9$$

This value is closer to the value of μ but is still not a very good estimate of the population mean. The problem here is that the population of x values exhibits a lot of variability (the largest value is $x = 7{,}767{,}421$ for Los Angeles County, California, and the smallest value is $x = 89$ for Loving County, Texas, which evidently few people love). Therefore, it is difficult for a sample of 15 observations, let alone just 5, to be reasonably representative of the population. In Chapter 9, you will see how variability is taken into account when deciding on the sample size required to accurately estimate a population mean.

One potential drawback to the mean as a measure of center for a data set is that its value can be greatly affected by the presence of even a single “outlier” (an unusually large or small observation) in the data set.

Example 4.3 Number of Visits to a Class Website

Understand the context ➤

● Forty students were enrolled in a section of a general education course in statistical reasoning. The instructor made course materials, grades, and lecture notes available to students on a class website, and course management software kept track of how often each student accessed any of the web pages on the class site.

One month after the course began, the instructor requested a report that indicated how many times each student had accessed a web page on the class site. The 40 observations were:

Consider the data ➤

20	37	4	20	0	84	14	36	5	331	19	0
0	22	3	13	14	36	4	0	18	8	0	26
4	0	5	23	19	7	12	8	13	16	21	7
13	12	8	42								

Interpret the results ➤

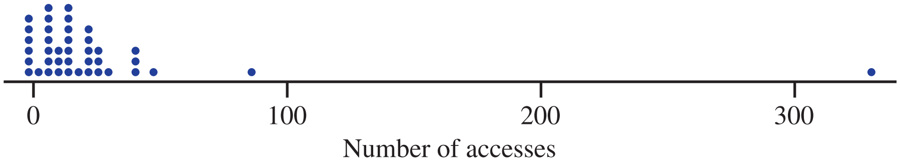
● **Data set available online**

The sample mean for this data set is $\bar{x} = 23.10$. Figure 4.3 is a Minitab dotplot of the data. Many would argue that 23.10 is not a very representative value for this sample, because 23.10

Copyright 2025 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s). Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

is larger than most of the observations in the data set. Notice that only 7 of 40 observations, or 17.5%, are larger than 23.10. The two outlying values of 84 and 331 (no, that was *not* a typo!) have a substantial impact on the value of $\bar{x}$.

Figure 4.3
A Minitab dotplot of the data in Example 4.3.



We now turn our attention to a measure of center that is not as sensitive to unusual values in a data set—the median.

The Median

The median strip of a highway divides the highway in half, and the median of a numerical data set does the same thing for a data set. Once the data values have been listed in order from smallest to largest, the **median** is the middle value in the list, and it divides the list into two equal parts.

Depending on whether the sample size n is even or odd, the process for determining the median is slightly different. When n is an odd number (say, 5), the sample median is the single middle value. But when n is even (say, 6), there are two middle values in the ordered list, and you average these two middle values to obtain the sample median.

Definition

Sample median: The **sample median** is obtained by first ordering the n observations from smallest to largest (with any repeated values included, so that every sample observation appears in the ordered list). Then

$$\text{sample median} = \begin{cases} \text{the single middle value if } n \text{ is odd} \\ \text{the average of the middle two values if } n \text{ is even} \end{cases}$$

Example 4.4 Website Data Revised

The sample size for the website access data of Example 4.3 was $n = 40$, an even number. The median is the average of the 20th and 21st values (the middle two) in the ordered list of the data. Arranging the data in order from smallest to largest produces the following ordered list (with the two middle values highlighted):

0	0	0	0	0	0	3	4	4	4	5	5
7	7	8	8	8	12	12	**13**	**13**	13	14	14
16	18	19	19	20	20	21	22	23	26	36	36
37	42	84	331								

The median can now be determined by averaging the two middle values:

Do the work ➤

$$\textit{median} = \frac{13 + 13}{2} = 13$$

Interpret the results ➤ Looking at the dotplot (Figure 4.3), this value appears to be a more typical value for the data set than the sample mean of 23.10.

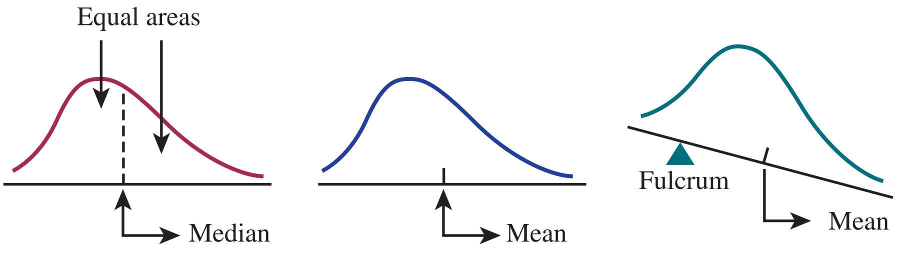
The sample mean can be sensitive to even a single value that lies far above or below the rest of the data. The value of the mean is pulled out toward such outlying values. The median, on the other hand, is quite *in*sensitive to outliers. For example, the largest sample observation (331) in Example 4.4 can be increased by any amount without changing the value of the median. Similarly, an increase in the second or third largest observations does not affect the median, nor would a decrease in several of the smallest observations.

This stability of the median is what justifies its use as a measure of center in some situations. For example, the article **"Average Credit Card Debt in America" (March 8, 2023, moneygeek.com)** reported that the mean credit card debt for Americans less than 35 years of age was $3700, whereas the median credit card debt was only $1900. In this case, a small percentage of Americans under age 35 with unusually high credit card debt results in a mean that may not be representative of a typical credit card debt for people in this age group.

Comparing the Mean and the Median

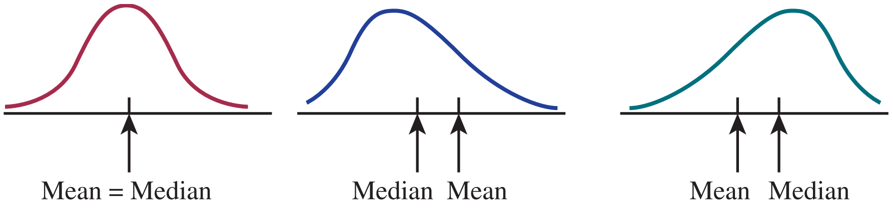
Figure 4.4 shows several smoothed histograms that might represent either a distribution of sample values or a population distribution. The median is the value on the measurement axis that separates the smoothed histogram into two parts, with half (50%) of the area under each part of the curve. The mean is a bit harder to visualize. If the histogram were balanced on a triangle (a fulcrum), it would tilt unless the triangle was positioned at the mean. The mean is the balance point for the distribution.

Figure 4.4
The mean and the median.



When the histogram is symmetric, the point of symmetry is both the dividing point for equal areas and the balance point, and the mean and the median are equal. However, when the histogram is unimodal (single-peaked) with a longer upper tail (positively skewed), the outlying values in the upper tail pull the mean up, so it generally lies above the median. For example, an unusually high exam score raises the mean but does not affect the median. Similarly, when a unimodal histogram is negatively skewed, the mean is generally smaller than the median (see Figure 4.5).

Figure 4.5
Relationship between the mean and the median.



Example 4.5 **Baseball Salaries**

Understand the context ➤

USA Today publishes professional baseball players' salaries at the beginning of each season. The following are the 2021 salaries for the Washington Nationals players in millions of dollars **(databases.usatoday.com/mlb-salaries/)**.

Copyright 2025 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s). Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

Consider the data ➤

Player	2021Salary (millions of dollars)	Player	2021Salary (millions of dollars)	Player	2021Salary (millions of dollars)
Scherzer	34.5	Gomes	6.0	Robles	0.6
Strasburg	32.0	Hudson	6.0	Suero	0.6
Corbin	24.4	Lester	4.8	Fedde	0.6
Turner	13.0	Ross	1.5	Rainey	0.6
Hand	10.1	Avila	1.5	Voth	0.6
Soto	8.5	Mercer	1.0	Finnegan	0.6
Harris	8.0	Harrison	1.0	Stevenson	0.6
Castro	7.0	Zimmerman	1.0	Clay	0.6
Schwarber	7.0	Avilan	1.0	Hernandez	0.6
Bell	6.4	Perez	0.8	Barrera	0.6

A dotplot of these data is shown in Figure 4.6.

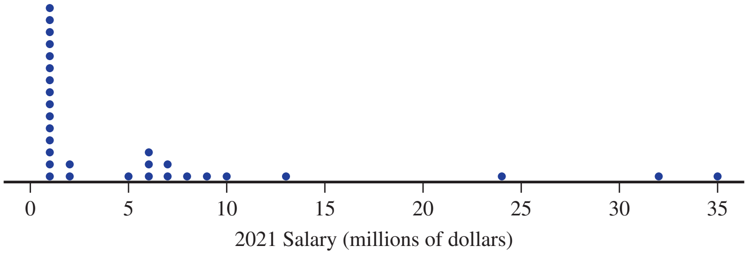


Figure 4.6
Dotplot for the Washington Nationals 2021 salary data.

What Do You Think?

How do you think the three unusually large salaries in the data set would affect the mean?

Because the data distribution is positively skewed and there are outliers, you would expect the mean to be greater than the median.

For this data set, the mean is

$$\bar{x} = \frac{34.5 + \cdots + 0.6}{30} = 6.05 \text{ million}$$

The median is the average of the middle two observations in the ordered list of salaries:

$$\text{median} = \frac{1.5 + 1.0}{2} = 1.25 \text{ million}$$

Notice that the median is smaller than the mean, and it is probably a better description of a typical salary for this group.

Categorical Data

The natural numerical summary quantities for a categorical data set are the relative frequencies for the different possible categories. Each relative frequency is the proportion (fraction) of responses in the corresponding category. Often there are only two possible responses (a dichotomy)—for example, does or does not have a driver's license, did or did not vote in the last election. It is convenient in such situations to label one of the two possible responses S (for success) and the other F (for failure). As long as further analysis is consistent with the labeling, it does not matter which category is assigned the S label. When the data set is a sample, the fraction of S's in the sample is called the **sample proportion of successes**.

Copyright 2025 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s). Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

Definition

Sample proportion of successes: The **sample proportion of successes**, denoted by $\hat{p}$, is

$$\hat{p} = \text{sample proportion of successes} = \frac{\text{number of S's in the sample}}{n}$$

where S is the label used for the response designated as success.

Example 4.6 Can You Hear Me Now?

Understand the context ➤ It is not uncommon for a cell phone user to complain about the quality of their service provider. Suppose that each person in a sample of $n = 15$ cell phone users is asked if they are satisfied with their cell phone service. Each response is classified as S (satisfied) or F (not satisfied). The resulting data are

Consider the data ➤

S F S S S F F S S F

S S S F F

Do the work ➤ This sample contains nine S's and $n = 15$, so

$$\hat{p} = \frac{9}{15} = 0.60$$

Interpret the results ➤ This means that 60% of the sample responses are S's. Of those surveyed, 60% are satisfied with their cell phone service.

The letter p is used to denote the **population proportion of S's**.* You will see later how the value of $\hat{p}$ from a particular sample can be used to draw conclusions about the population proportion p.

---

*Note that this is one situation in which we will not use a Greek letter to denote a population characteristic. Some statistics books use the symbol π for the population proportion and p for the sample proportion. We will not use π in this context so that there is no confusion with the mathematical constant $\pi = 3.14\ldots$.

Exercises 4.1–4.16

● Data set available online

4.1 The following are the prices (in dollars) of the six all-terrain truck tires rated most highly by *Consumer Reports* in 2018 **(consumerreports.org)**:

159.00 199.00 157.00 127.65 123.99 126.00

a. Calculate the values of the mean and median.
b. Why are these values so different?
c. Which of the two—mean or median—appears to be better as a description of a typical value for this data set? (Hint: See Example 4.5.)

4.2 ● The article **"Caffeine Content of Drinks" (caffeineinformer.com/the-caffeine-database)** gave the following data on caffeine concentration (mg/ounce) for eight top-selling energy drinks:

Energy Drink	Caffeine Concentration (mg/oz)
Red Bull	9.5
Monster	10.0
Rockstar	10.0
Full Throttle	10.0
No Fear	11.4
Amp	8.9
SoBe Adrenaline Rush	9.5
Tab Energy	9.1

Copyright 2025 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s). Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

a. What is the value of the mean caffeine concentration for this set of top-selling energy drinks?

b. Coca-Cola has 2.9 mg/ounce of caffeine and Pepsi Cola has 3.2 mg/ounce of caffeine. Write a sentence explaining how the caffeine concentration of top-selling energy drinks compares to that of these colas.

4.3 ● **Consumer Reports Health (consumerreports.org/health)** reported the accompanying caffeine concentration (mg/cup) for 12 brands of coffee:

Coffee Brand	Caffeine Concentration (mg/cup)
Eight O'Clock	140
Caribou	195
Kickapoo	155
Starbucks	115
Bucks Country Coffee Co.	195
Archer Farms	180
Gloria Jean's Coffees	110
Chock Full o'Nuts	110
Peet's Coffee	130
Maxwell House	55
Folgers	60
Millstone	60

Use at least one measure of center to compare caffeine concentration for these coffees with that of the energy drinks of the previous exercise. (Note: 1 cup = 8 ounces)

4.4 ● The author of the paper **"Gambling with Your Health: Bacterial Contamination on Casino Gambling Chips" (*Journal of Environmental Health* [2019]: 8–14)** looked at bacteria on 13 chips collected from several different casinos. They swabbed the front of each chip and then the swabs were streaked across a petri dish. After 48 hours, the number of *E. coli* bacteria found were counted for each chip.

15 14 4 4 4 4 0 19 7 1 11 1 1

Would you use the mean or the median to describe the center of this data set? Explain your choice.

4.5 ● *Consumer Reports* surveyed subscribers and asked them to rate their Internet service providers. The survey responses were then used to determine an overall satisfaction score. The accompanying data are the overall satisfaction ratings for 26 Internet service providers **(*Consumer Reports,* August 2021, p. 35).**

85 82 71 70 70 70 70 68 67 67 64 64 63
63 63 62 62 61 61 60 58 58 57 56 51 46

a. Display these data using a dotplot. Comment on any unusual features of the plot.

b. Calculate the mean and median satisfaction score for this sample.

c. The values of the mean and the median for this data set are similar. What aspect of the distribution of satisfaction score—as pictured in the dotplot from Part (a)—provides an explanation for why the values of the mean and median are similar? (Hint: See the discussion of Figure 4.4.)

4.6 The state of California defines family income groups in terms of median county income as follows:

Extremely low income: below 30% of county median income

Very Low income: between 30% and 50% of county median income

Low income: between 50% and 80% of county median income

Moderate income: between 80% and 120% of county median income

For San Luis Obispo county, the median household income in 2021 was $82,514 **(census.gov)**.

a. Interpret the value of the median household income in 2021 for San Luis Obispo County.

b. Each of the following statements is incorrect. For each statement, use the given information to explain why it is incorrect.

Statement 1: 30% of the households in San Luis Obispo County would be classified as extremely low income.

Statement 2: More than 50% of the households in San Luis Obispo County would be classified as extremely low income or very low income.

Statement 3: There cannot be any households in San Luis Obispo County that would be classified as having an income that was greater than those in the moderate-income category.

4.7 ● The report titled **"Top Ten Newspaper Circulations: Biggest Print Titles Have Lost 30% of Sales Since 2016 Election" (PressGazette, October 24, 2020)** includes the weekday circulation numbers for the top 10 newspapers in the United States. Here are the data for the first quarter of 2020:

994,600 486,579 410,562 206,824 193,015

162,478 137,590 130,357 123,905 123,881

a. Calculate and interpret the value of the median of this data set.

b. Explain why the median is preferable to the mean for describing center for this data set.

c. Explain why it would be unreasonable to generalize from this sample of 10 newspapers to the population of all daily newspapers in the United States.

4.8 Each student in a sample of 20 seniors at a particular university was asked if they were registered to vote. With R denoting registered and N denoting not registered, the sample data are:

R R N N N R N R N R
N R N R R R N R R R

a. If being registered to vote is considered a "success," what is the value of the proportion of successes for this sample?
b. When would it be reasonable to generalize from this sample to the population of all seniors at this university?

4.9 The report **"The Plastic Waste Makers Index" (minderoo.org/plastic-waste-makers-index/)** gives data on the single-use plastic waste generation per capita (in kg per person) for countries around the world. The accompanying data are the values for the 11 countries with the highest single-use waste generation.

76 59 56 55 55 55 55 53 53 52 51

a. What summary measures would you use to describe center for this data set? Explain your choice.
b. Explain why it would not be reasonable to draw conclusions about per-capita single-use plastic waste generation for all countries in the world based on the given sample.

4.10 The ministry of **Health and Long-Term Care in Ontario, Canada,** publishes information on its website **(health.gov.on.ca)** on the time that patients must wait for various medical procedures. For two cardiac procedures, the following information was provided:

	Number of Completed Procedures	Median Wait Time (days)	Mean Wait Time (days)	90% Completed Within (days)
Angioplasty	847	14	18	39
Bypass surgery	539	13	19	42

The median wait time for angioplasty is greater than the median wait time for bypass surgery but the mean wait time is shorter for angioplasty than for bypass surgery. What does this suggest about the distribution of wait times for these two procedures?

4.11 ● In the report **"Your Driving Costs 2020,"** the Automobile Club of America estimated the cost of operating a car for different types of new 2020 model vehicles (**ace.aaa.com**). The following table gives the estimated operating cost per year (in dollars).

Vehicle Type	Estimated Total Cost per Year (dollars)
Small Sedan	7,516
Medium Sedan	8,734
Large Sedan	11,182
Small SUV	8,362
Medium SUV	10,036
Minivan	10,101
Pickup	11,308
Hybrid	8,148
Electric	9,119

a. Calculate the median estimated cost for these nine types of vehicles.
b. Explain why the median might be preferable to the mean for describing center for this data set.
c. Do you think that the mean for this data set would be greater than, about the same as, or less than the value of the median you calculated in Part (a)? Explain your choice.

4.12 Consider the following statement: More than 65% of the residents of Los Angeles earn less than the average wage for that city. Could this statement be correct? If so, how? If not, why not?

4.13 ● The accompanying data on the number of medical errors involving prescriptions in a 24-hour period for each of 10 emergency rooms is from a paper that appeared in the ***Journal of Critical Care*** **([2019]: 258–263)**.

1 0 10 1 1 2 2 18 6 8

a. Construct a dotplot of the data. Will the mean or the median be larger for this data set? Explain your reasoning.
b. Calculate the values of the mean and median.
c. By how much could the largest value in the data set be increased without affecting the value of the sample median? By how much could this value be decreased without affecting the value of the median?

4.14 Suppose that 10 patients with meningitis received treatment with large doses of penicillin. Three days later, temperatures were recorded, and the treatment was considered successful if there had been a reduction in a patient's temperature. Denoting success by S and failure by F, the 10 observations are

S S F S S S F F S S

a. What is the value of the sample proportion of successes?

b. Replace each S with a 1 and each F with a 0. Then calculate $\bar{x}$ for this numerically coded sample. How does $\bar{x}$ compare to $\hat{p}$?

c. Suppose that it is decided to include 15 more patients in the study. How many of these would have to be S's to give $\hat{p} = 0.80$ for the entire sample of 25 patients?

4.15 A study of the lifetime (in hours) for a certain brand of light bulb involved putting 10 light bulbs into operation and observing them for 1000 hours. Eight of the light bulbs failed during that period, and those lifetimes were recorded. The lifetimes of the two light bulbs still functioning after 1000 hours were recorded as 1000+. The resulting sample observations were

480 790 1000+ 350 920 860 570 1000+
170 290

Which of the measures of center discussed in this section can be calculated, and what are the values of those measures?

4.16 An instructor has graded 19 exam papers submitted by students in a class of 20 students, and the average so far is 70. (The maximum possible score is 100.) How high would the score on the last paper have to be to raise the class average by 1 point? By 2 points?

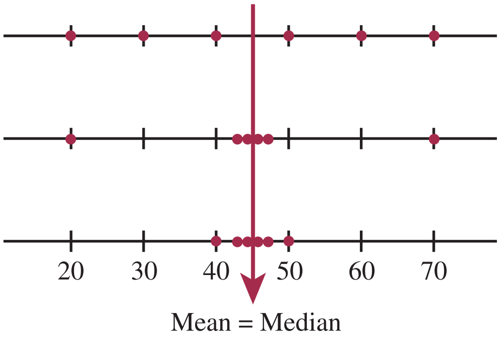
Section 4.2 Describing Variability in a Data Set

Reporting a measure of center gives only partial information about a data set. It is also important to describe how much the observations differ from one another. The three different samples displayed in Figure 4.7 all have mean = median = 45. There is a lot of variability in the first sample compared to the third sample. The second sample shows less variability than the first and more variability than the third. Most of the variability in the second sample is due to the two extreme values being so far from the center.

Figure 4.7
Three samples with the same center and different amounts of variability.

Sample

1. 20, 40, 50, 30, 60, 70
2. 47, 43, 44, 46, 20, 70
3. 44, 43, 40, 50, 47, 46



The simplest numerical measure of variability is the range.

Definition

Range: The **range** of a data set is defined as Range = Largest Value − Smallest Value

In general, more variability will result in a larger range. However, variability is a characteristic of the entire data set, and each observation contributes to variability. The first two samples plotted in Figure 4.7 both have a range of 70 − 20 = 50, but there is less variability in the second sample. Because it is calculated using only the largest and smallest values in the data set, the range is not usually the best measure of variability.

Deviations from the Mean

The most widely used measures of variability describe the extent to which the sample observations deviate from the sample mean $\bar{x}$. Subtracting $\bar{x}$ from each observation gives a set of **deviations from the mean**.

Definition

Deviations from the mean: The ***n*** **deviations from the sample mean** are the differences

$$(x_1 - \bar{x}), (x_2 - \bar{x}), \ldots, (x_n - \bar{x})$$

Notice that a deviation will be positive if the corresponding x value is greater than $\bar{x}$ and negative if the x value is less than $\bar{x}$.

Example 4.7 The Big Mac Index

Understand the context ➤

● McDonald's fast-food restaurants are found in many countries around the world. Table 4.3 shows the cost of a Big Mac for 9 Central and South American countries in June 2021 (converted to U.S. dollars) from the article **"The Big Mac Index" (economist.com/big-mac-index, 2021).**

Consider the data ➤

Table 4.3 Big Mac Prices for Nine Central and South American Countries

Country	June 2021Big Mac Price in U.S. Dollars
Argentina	3.75
Brazil	3.98
Chile	4.09
Colombia	3.74
Costa Rica	3.83
Guatemala	3.20
Honduras	3.61
Nicaragua	3.56
Peru	3.29

Notice that there is quite a bit of variability in the Big Mac prices. For this data set, $\Sigma x = 33.05$ and $\bar{x} = \$3.67$. Table 4.4 displays the data along with the corresponding deviations, formed by subtracting $\bar{x} = \$3.67$ from each observation. Five of the deviations are positive because five of the observations are larger than $\bar{x}$. Some of the deviations are quite large in magnitude (-0.47 and -0.38, for example), indicating observations that are far from the sample mean.

Table 4.4 Deviations from the Mean for the Big Mac Data

June 2021Big Mac Price in U.S. Dollars	Deviation from the Mean ($x - \bar{x}$)
3.75	0.08
3.98	0.31
4.09	0.42
3.74	0.07
3.83	0.16
3.20	−0.47
3.61	−0.06
3.56	−0.11
3.29	−0.38

● **Data set available online**

In general, the greater the amount of variability in the sample data, the larger the magnitudes (ignoring the signs) of the deviations. Next, consider how to combine the deviations into a single numerical measure of variability. A first thought might be to calculate the average deviation, by adding the deviations together and then dividing by n. This does not work, though, because when you calculate the sum of the deviations, denoted by $\Sigma(x - \bar{x})$, negative and positive deviations offset each other. In fact, except for small differences due to rounding, the sum of the deviations from the mean is always equal to 0. Notice that the value of the sum of the 9 deviations in Example 4.7 is $\Sigma(x - \bar{x}) = 0.02$.

Except for the effects of rounding in calculating the deviations, it is always true that

$$\Sigma(x - \bar{x}) = 0$$

Since this sum is zero, the average deviation is always zero and so it is not useful as a measure of variability.

The Variance and Standard Deviation

One way to prevent negative and positive deviations from offsetting one another is to square them before combining. Then deviations with opposite signs but with the same magnitude, such as $+2$ and -2, make identical contributions to a measure of variability. The squared deviations are $(x_1 - \bar{x})^2, (x_2 - \bar{x})^2, \ldots, (x_n - \bar{x})^2$ and the sum of the squared deviations is

$$(x_1 - \bar{x})^2 + (x_2 - \bar{x})^2 + \cdots + (x_n - \bar{x})^2 = \Sigma(x - \bar{x})^2$$

Dividing this sum by the sample size n gives the average squared deviation. Although this seems to be a reasonable measure of variability, a divisor slightly smaller than n is used. (The reason for this will be explained later in this section and in Chapter 9.)

Definitions

Sample variance: The **sample variance**, denoted by s^2, is the sum of squared deviations from the mean divided by $n - 1$. That is,

$$s^2 = \frac{\Sigma(x - \bar{x})^2}{n - 1}$$

Sample standard deviation: The **sample standard deviation** is the positive square root of the sample variance and is denoted by $\boldsymbol{s}$.

A large amount of variability in a sample is indicated by a relatively large value of s^2 or s, whereas a value of s^2 or s close to zero indicates a smaller amount of variability. People find the standard deviation to be a more natural measure of variability than the variance because the standard deviation is expressed in the same units as the original data values. For example, if the observations in the data set are the prices of a Big Mac in dollars, the mean and the deviations from the mean are also in dollars. But when the deviations are squared and then combined to calculate the value of the variance, the units are dollars squared—not something that is familiar to most people! This makes it difficult to interpret the value of the variance and to decide whether the variance is large or small. Taking the square root of the variance to obtain the standard deviation results in a measure of variability that is in the same units as the original data values, making it easier to interpret.

Example 4.8 Big Mac Revisited

Let's use the Big Mac data and the deviations from the mean calculated in Example 4.7 to calculate the values of the sample variance and the sample standard deviation. Table 4.5 shows the original observations and the deviations from the mean, along with the squared deviations.

Table 4.5 Deviations and Squared Deviations for the Big Mac Data

June 2021Big Mac Price in U.S. Dollars	Deviation from the Mean $(x - \bar{x})$	Squared Deviation from the Mean $(\bar{x} - \bar{x})^2$
3.75	0.08	0.0064
3.98	0.31	0.0961
4.09	0.42	0.1764
3.74	0.07	0.0049
3.83	0.16	0.0256
3.20	−0.47	0.2209
3.61	−0.06	0.0036
3.56	−0.11	0.0121
3.29	−0.38	0.1444
		$\Sigma(x - \bar{x})^2 = 0.6904$

Combining the squared deviations to calculate the values of s^2 and s gives

$$s^2 = \frac{\Sigma(x - \bar{x})^2}{n - 1} = \frac{0.6904}{9 - 1} = \frac{0.6904}{8} = 0.0863$$

$$s = \sqrt{0.0863} = 0.2938$$

The calculation of s^2 can be a bit tedious, especially if the sample size is large. Fortunately, many calculators and computer software packages compute the variance and standard deviation. One commonly used statistical computer package is Minitab. The output resulting from using the Minitab Describe command with the Big Mac data follows. Minitab gives a variety of numerical descriptive measures, including the mean, the median, and the standard deviation.

```
Statistics

Variable      N  N*    Mean  SE Mean   StDev  Minimum      Q1  Median      Q3  Maximum

2021 Big      9   0  3.6722   0.0979  0.2938   3.2000  3.4250  3.7400  3.9050   4.0900
Mac Price
in U.S.
Dollars
```

The standard deviation can be informally interpreted as the size of a "typical" or "representative" deviation from the mean. In Example 4.8, a typical deviation from $\bar{x}$ is about 0.2938. Some observations are closer to $\bar{x}$ than 0.2938 and others are farther away.

The standard deviation $s = 0.2938$ in Example 4.8 was calculated without saying whether this value indicated a large or a small amount of variability. At this point, it is better to use s for comparative purposes than for an absolute assessment of variability. If Big Mac prices for a larger group of countries resulted in a standard deviation of $s = 1.153$ then you would conclude that the original sample has much less variability than the data set consisting of the larger group of countries.

There are measures of variability for the entire population that are analogous to s^2 and s for a sample. These measures are called the **population variance** and the **population standard deviation** and are denoted by $\boldsymbol{\sigma}^2$ and $\boldsymbol{\sigma}$, respectively. (Lowercase Greek letters are used for population characteristics.)

Notation

s^2	sample variance
σ^2	population variance
s	sample standard deviation
σ	population standard deviation

In many statistical procedures, you would like to use the value of σ, but it is not usually known. Therefore, you must estimate σ using a value calculated from the sample. The divisor $(n - 1)$ is used in calculating s^2 rather than n because, on average, the resulting value tends to be a bit closer to the value of σ^2. More will be said about this in Chapter 9.

An alternative rationale for using $(n - 1)$ is based on the property $\Sigma(x - \bar{x}) = 0$. Suppose that $(n = 5)$ and that four of the deviations are

$$(x_1 - \bar{x}) = -4 \qquad (x_2 - \bar{x}) = 6 \qquad (x_3 - \bar{x}) = 1 \qquad (x_5 - \bar{x}) = -8$$

Then, because the sum of these four deviations is -5, the remaining deviation must be $(x_4 - \bar{x}) = 5$ (so that the sum of all five deviations is zero). More generally, once any $(n - 1)$ of the deviations are known, the value of the remaining deviation is determined. The n deviations actually contain only $(n - 1)$ independent pieces of information about variability. Statisticians express this by saying that s^2 and s are based on $(n - 1)$ *degrees of freedom* (df).

The Interquartile Range

As with $\bar{x}$, the value of s can be greatly affected by the presence of even a single unusually small or large observation. The **interquartile range** is a measure of variability that is resistant to the effects of outliers. It is based on quantities called quartiles. The **lower quartile** separates the smallest 25% of the data set from the largest 75%, and the **upper quartile** separates the largest 25% from the smallest 75%. The middle quartile is the median, and it separates the lower 50% from the upper 50%. Figure 4.8 illustrates the locations of these quartiles for a smoothed histogram.

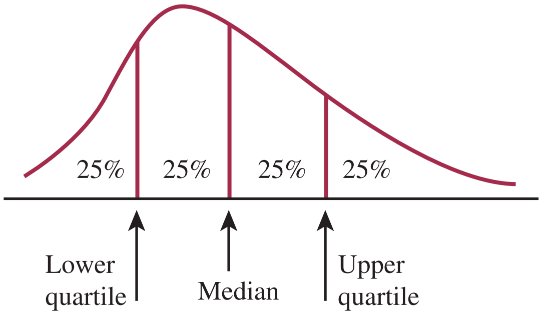


Figure 4.8
The quartiles for a smoothed histogram.

The quartiles for sample data are obtained by dividing the n ordered observations into a lower half and an upper half. If n is odd, the median is excluded from both halves. The upper and lower quartiles are then the medians of the two halves. (Note: The median is only temporarily excluded for the purpose of calculating quartiles. It is not deleted from the data set.)

Lower quartile: Median of the lower half of the sample

Upper quartile: Median of the upper half of the sample

(If n is odd, the median of the entire sample is excluded from both halves when calculating quartiles.)

Interquartile range (iqr): A measure of variability that is not as sensitive to the presence of outliers as the standard deviation. The iqr is calculated as

$$\textbf{iqr} = \textbf{upper quartile} - \textbf{lower quartile}$$

Note: There are several other sensible ways to define quartiles. Some calculators and software packages use alternative definitions.

The resistant nature of the interquartile range follows from the fact that up to 25% of the smallest sample observations and up to 25% of the largest sample observations can be made more extreme without affecting the value of the interquartile range.

Example 4.9 Beating That High Score

Understand the context ➤

● The authors of the paper **"Striatal Volume Predicts Level of Video Game Skill Acquisition" (*Cerebral Cortex* [2010]: 2522–2530)** studied a number of factors that affect performance on a complex video game. One of the factors investigated was practice strategy. Forty college students who all reported that they played video games less than 3 hours per week over the 2 years prior to the study and who had never played the game Space Fortress were assigned at random to one of two practice strategies. In the Space Fortress game, points are awarded for control, velocity, and speed.

Each person completed 20 two-hour practice sessions. Those in one group, the fixed priority group, were told to work on improving their total score at each practice session. Those in the other group, the variable priority group, were told to focus on a particular aspect of the game, such as improving speed score, in each practice session, and the focus changed from one practice session to another. The investigators were interested in whether practice strategy makes a difference. They measured the improvement in total score from the first day of the study to the day when the last practice session was completed.

Improvement scores (approximate values read from a graph that appears in the paper) for the 20 people in the variable priority practice strategy group were ordered from smallest to largest and given here.

Consider the data ➤

1200	1300	2300	3200	3300	3800	4000	4100	4300	4800
5500	5700	5700	5800	6000	6300	6800	6800	6900	7700

The median is the average of the middle two observations, so

$$\text{median} = \frac{4800 + 5500}{2} = 5150$$

The lower half of the data set is

1200 1300 2300 3200 3300 3800 4000 4100 4300 4800

so the lower quartile is the median of the lower half

$$\text{lower quartile} = \frac{3300 + 3800}{2} = 3550$$

The upper half of the data set is

5500 5700 5700 5800 6000 6300 6800 6800 6900 7700

● **Data set available online**

so the upper quartile is the median of the upper half

$$\text{upper quartile} = \frac{6000 + 6300}{2} = 6150$$

$$\text{lower quartile} = 3550$$
$$\text{upper quartile} = 6150$$
$$\text{iqr} = 6150 - 3550 = 2660$$

The sample mean and standard deviation for this data set are 4775 and 1867, respectively. If you were to change the two largest values from 6900 and 7700 to 10,900 and 11,700 (so that they still remain the two largest values), the median and interquartile range would not be affected, but the mean and the standard deviation would change to 7500 and 2390, respectively. The value of the interquartile range is not affected by extreme values in the data set.

Example 4.10 Some Strange Things Can Happen with Quartiles…

● The Chain Reaction Scorecard rates the top 20 fast-food and fast-casual restaurant chains in the United States that serve beef based on the policies and actions related to antibiotic use in their beef supplies. The report **"Chain Reaction VI: How Top Restaurants Rate on Reducing Antibiotic Use in Their Beef Supply Chains" (advocacy.consumerreports.org, July 2021)** includes the following ratings for 2021. This data set is interesting because some data values appear multiple times. For example, there are 12 observations that all have the value 0.

Because the distribution is skewed and has outliers, the median and interquartile range would be an appropriate choice for describing center and variability. The ordered data values are

0 0 0 0 0 0 0 0 0 **0**
0 0 19 19 28 41 42 48 85 94

The median is the average of the two middle values. Here, both middle values are 0, and so the median is also 0.

Dividing the data set into the lower and upper halves and then finding the median of each half leads to the following quartiles.

Lower half

0 0 0 0 **0** **0** 0 0 0 0

$$\text{lower quartile} = 0$$

Upper half

0 0 19 19 **28** **41** 42 48 85 94

$$\text{upper quartile} = \frac{28 + 41}{2} = 34.5$$

You can now calculate the value of the interquartile range.

$$\text{iqr} = \text{upper quartile} - \text{lower quartile} = 34.5 - 0 = 34.5$$

Notice here that the lower quartile and the median are both 0. This is because more than half of the observations in the data set have a value of 0. For this data set, the smallest 25% are all 0, and even the next 25% in the ordered list are all 0. One way to see this is to take another look at the ordered data list:

$$\underbrace{0\ 0\ 0\ 0\ 0}_{\text{25\% of observations}}\quad \underbrace{\underbrace{0\ 0\ 0\ 0\ 0}_{\text{25\% of observations}}\quad \underbrace{0\ 0\ 19\ 19\ 28}_{\text{25\% of observations}}}_{\text{middle 50\%}}\quad \underbrace{41\ 42\ 48\ 85\ 94}_{\text{25\% of observations}}$$

● **Data set available online**

The value of the interquartile range indicates that there is quite a bit of variability in the middle half of the data distribution.

Exercises 4.17–4.33

● **Data set available online**

4.17 ● The following data are costs (in cents) per ounce for nine different brands of sliced Swiss cheese (consumerreports.org):

29 62 37 41 70 82 47 52 49

a. Calculate the variance and standard deviation for this data set. (Hint: See Example 4.8.)

b. If a very expensive cheese with a cost per ounce of $1.50 (150 cents) was added to the data set, how would the values of the mean and standard deviation change?

4.18 ● Cost per serving (in cents) for six high-fiber cereals rated very good and for nine high-fiber cereals rated good by *Consumer Reports* are shown below. Write a few sentences describing how these two data sets differ with respect to center and variability. Use summary statistics to support your statements.

Cereals Rated Very Good

46 49 62 41 19 77

Cereals Rated Good

71 30 53 53 67 43 48 28 54

4.19 ● In an experiment to assess the effect of listening to audiobooks while driving, participants were asked to drive down a straight road in a driving simulator. The accompanying data on time (in milliseconds) to react when a pedestrian walked into the street for 10 drivers listening to an audiobook are consistent with summary statistics and graphs that appeared in the paper "Good Distractions: Testing the Effect of Listening to an Audiobook on Driving Performance in Simple and Complex Road Environments" (*Accident Analysis and Prevention* [2018]: 202–209). Calculate and interpret the values of the median and the interquartile range for this data set.

1014 1009 1052 982 931
1020 1069 1011 860 1108

4.20 ● The 2015 Urban Mobility Scorecard (Texas A&M Transportation Institute, mobility.tamu.edu) includes data on the estimated cost (in millions of dollars) resulting from traffic congestion for different urban areas. The following data are for the 13 largest U.S. urban areas that had a population of over 3 million.

Urban Area	Total Cost (millions of dollars)
New York	15
Los Angeles	13
Chicago	7
Washington, DC	5
Houston	5
Dallas/Fort Worth	4
Detroit	4
Miami	4
Phoenix	4
Philadelphia	4
San Francisco	3
Boston	3
Atlanta	3

Calculate and interpret the values of the quartiles and the interquartile range.

4.21 ● The accompanying data are from the paper "Ecological Sexual Dimorphism and Environmental Variability within a Community of Antarctic Penguins" (journals.plos.org, 2014). They represent the body mass (in grams) for a sample of 10 female chinstrap penguins. Calculate the mean and standard deviation for this data set.

3400 3200 3850 3600 3700 3400 3250 3600 3400 3700

4.22 ● The paper referenced in the previous exercise also gave body mass for a sample of male chinstrap penguins.

4050 4300 3800 4800 3775 3950 3650 4100 4500 4400

a. Use the data given in this exercise and the data given in the previous exercise to construct dotplots that will allow you to compare the body mass measurements for the male and the female chinstrap penguins.

b. Based on the dotplots, do you think the mean body mass for the sample of males will be greater than, about the same as, or less than the mean that you calculated in the previous exercise? Explain your thinking.

c. Based on the dotplots, do you think the standard deviation of the body masses for the sample of males will be greater than, about the same as, or less than the standard deviation that you calculated in the previous exercise? Explain your thinking.

Copyright 2025 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s). Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

d. Calculate the values of the mean and standard deviation for the sample of male body mass measurements. Are these values consistent with your answers to Parts (b) and (c)?

e. Describe how the body mass distributions differ for female and male chinstrap penguins.

4.23 ● The price (in dollars) of the eight smart phones that were rated highest by *Consumer Reports* in 2018 **(consumerreports.org)** were

730 850 830 800 700 1000 950 520

a. Calculate the values of the variance and the standard deviation.

b. The standard deviation is quite large. What does that tell you about the prices of these highly rated smart phones?

4.24 ● In addition to the prices of the highly rated smart phones given in the previous exercise, *Consumer Reports* also gave the prices of the seven smart phones that received the lowest ratings. Those prices (in dollars) were

110 130 100 145 200 180 135

Comment on how the highest rated smart phones and the lowest rated smart phones differ with respect to price and price variability.

4.25 In an experiment to assess the effect of listening to audiobooks while driving, participants were asked to drive down a straight road in a driving simulator. The accompanying data on time (in milliseconds) to react when a pedestrian walked into the street for 10 drivers listening to an audiobook are consistent with summary statistics and graphs that appeared in the paper **"Good Distractions: Testing the Effect of Listening to an Audiobook on Driving Performance in Simple and Complex Road Environments" (*Accident Analysis and Prevention* [2018]: 202–209)**. Calculate the variance and the standard deviation for this data set.

1014 1009 1052 982 931 1020 1069 1011 860 1108

4.26 ● The paper referenced in the previous exercise also gave summary statistics and graphs for the reaction time of drivers who were not listening to audiobooks. Data on reaction time (in milliseconds) consistent with those summary statistics for 10 drivers not listening to audiobooks are given here.

961 904 1010 976 1018 1041 1004 981 995 991

a. Use the data given in this exercise and the data given in the previous exercise to construct dotplots that would allow you to compare the reaction times for the two groups.

b. Based on the dot plots, do you think that the standard deviation of the reaction times for people who are not listening to audiobooks would be less than, about the same as, or greater than the standard deviation that you calculated in the previous exercise for drivers who were listening to audiobooks? Explain your thinking.

c. Calculate the standard deviation of the reaction times for the drivers who were not listening to audiobooks. Is the value of this standard deviation consistent with your answer in Part (b)?

d. Describe how the distributions of reaction time differ for drivers who are listening to audiobooks and those who are not.

4.27 ● Consider the following data on minutes of cell phone use in a 1-month period for a sample of 25 people:

189 0 189 177 106 201 0 212 0 306
0 0 59 224 0 189 142 83 71 165
236 0 142 236 130

a. Calculate the values of the quartiles and the interquartile range for this data set.

b. Explain why the lower quartile is equal to the minimum value for this data set. Will this be the case for every data set? Explain.

4.28 Give two sets of five numbers that have the same mean but different standard deviations, and give two sets of five numbers that have the same standard deviation but different means.

4.29 Morningstar is an investment research firm that publishes some online educational materials. The materials for an online course called **"Looking at Historical Risk" (morningstar.com/start-investing/classroom)** included the following paragraph referring to annual return (in percent) for investment funds:

> Using standard deviation as a measure of risk can have its drawbacks. It's possible to own a fund with a low standard deviation and still lose money. In reality, that's rare. Funds with modest standard deviations tend to lose less money over short time frames than those with high standard deviations. For example, the one-year average standard deviation among ultrashort-term bond funds, which are among the lowest-risk funds around (other than money market funds), is a mere 0.64%.

a. Explain why the standard deviation of percent return is a reasonable measure of unpredictability and why a smaller standard deviation for the percent return of an investment fund means less risk.

b. Explain how a fund with a small standard deviation can still lose money. (Hint: Think about the average percent return.)

4.30 ● The **U.S. Department of Transportation** reported the data in the accompanying table on the number of speed-related crash fatalities during holiday periods for the years from 1994 to 2003 **(*Traffic Safety Facts,* July 20, 2005)**.

a. Calculate the standard deviation for the New Year's Day data.

Data For Exercise 4.30

	Speed-Related Fatalities									
Holiday Period	**1994**	**1995**	**1996**	**1997**	**1998**	**1999**	**2000**	**2001**	**2002**	**2003**
New Year's Day	141	142	178	72	219	138	171	134	210	70
Memorial Day	193	178	185	197	138	183	156	190	188	181
July 4th	178	219	202	179	169	176	219	64	234	184
Labor Day	183	188	166	179	162	171	180	138	202	189
Thanksgiving	212	198	218	210	205	168	187	217	210	202
Christmas	152	129	66	183	134	193	155	210	60	198

b. Without calculating the standard deviation of the Memorial Day data, explain whether the standard deviation for the Memorial Day data would be larger or smaller than the standard deviation of the New Year's Day data.

c. Memorial Day and Labor Day are holidays that always occur on Monday and Thanksgiving always occurs on a Thursday, whereas New Year's Day, July 4th, and Christmas do not always fall on the same day of the week every year. Based on the given data, is there more or less variability in the speed-related crash fatality numbers from year to year for same day of the week holiday periods than for holidays that can occur on different days of the week? Support your answer with appropriate measures of variability.

4.31 **The Ministry of Health and Long-Term Care in Ontario, Canada,** publishes information on the time that patients must wait for various medical procedures on its website **(health.gov.on.ca)**. For two cardiac procedures, the following information was provided:

Procedure	**Number of Completed Procedures**	**Median Wait Time (days)**	**Mean Wait Time (days)**	**90% Completed Within (days)**
Angioplasty	847	14	18	39
Bypass surgery	539	13	19	42

a. Which of the following must be true for the lower quartile of the data set consisting of the 847 wait times for angioplasty?
- **i.** The lower quartile is less than 14.
- **ii.** The lower quartile is between 14 and 18.
- **iii.** The lower quartile is between 14 and 39.
- **iv.** The lower quartile is greater than 39.

b. Which of the following must be true for the upper quartile of the data set consisting of the 539 wait times for bypass surgery?
- **i.** The upper quartile is less than 13.
- **ii.** The upper quartile is between 13 and 19.
- **iii.** The upper quartile is between 13 and 42.
- **iv.** The upper quartile is greater than 42.

c. Which of the following must be true for the number of days for which only 5% of the bypass surgery wait times would be longer?
- **i.** It is less than 13.
- **ii.** It is between 13 and 19.
- **iii.** It is between 13 and 42.
- **iv.** It is greater than 42.

4.32 ● In 1997, a woman sued a computer keyboard manufacturer, charging that her repetitive stress injuries were caused by the keyboard ***(Genessey v. Digital Equipment Corporation)***. The jury awarded about $3.5 million for pain and suffering, but the court then set aside that award as being unreasonable compensation. In making this determination, the court identified a "normative" group of 27 similar cases and specified a reasonable award as one within 2 standard deviations of the mean of the awards in the 27 cases.

The 27 award amounts were (in thousands of dollars)

37	60	75	115	135	140	149	150
238	290	340	410	600	750	750	750
1050	1100	1139	1150	1200	1200	1250	1576
1700	1825	2000					

What is the maximum possible amount that could be awarded under the "2-standard deviations rule"?

4.33 ● The standard deviation alone does not measure relative variation. For example, a standard deviation of $1 would be considered large if it is describing the variability from store to store in the price of an ice cube tray. On the other hand, a standard deviation of $1 would be considered small if it is describing store-to-store variability in the price of a particular brand of freezer.

A quantity designed to give a relative measure of variability is the *coefficient of variation*. Denoted by CV, the coefficient of variation expresses the standard deviation as a percentage of the mean. It is defined by the formula $CV = 100\left(\frac{s}{\bar{x}}\right)$.

Consider two samples. Sample 1 gives the actual weight (in ounces) of the contents of cans of pet food labeled as having a net weight of 8 ounces. Sample 2 gives the actual weight (in pounds) of the contents of

Copyright 2025 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s). Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

bags of dry pet food labeled as having a net weight of 50 pounds. The weights for the two samples are

Sample 1	8.3	7.1	7.6	8.1	7.6
	8.3	8.2	7.7	7.7	7.5
Sample 2	52.3	50.6	52.1	48.4	48.8
	47.0	50.4	50.3	48.7	48.2

a. For each of the given samples, calculate the mean and the standard deviation.

b. Calculate the coefficient of variation for each sample. Do the results surprise you? Why or why not?

Section 4.3 Summarizing a Data Set: Boxplots

Sections 4.1 and 4.2 considered ways to describe the center and variability of a data set using numerical measures. It would be nice to have a method of summarizing data that provides more information than just reporting a measure of center and spread and yet less detail than a stem-and-leaf display or histogram. A **boxplot** is one way to do this. A boxplot is compact, yet it provides information about the center, variability, and symmetry or skewness of the data. Two types of boxplots are considered: the skeletal boxplot and the modified boxplot.

Construction of a Skeletal Boxplot

1. Draw a horizontal (or vertical) measurement scale.
2. Construct a rectangular box with a left (or lower) edge at the lower quartile and a right (or upper) edge at the upper quartile. This makes the box width equal to the iqr.
3. Draw a vertical (or horizontal) line segment inside the box at the location of the median.
4. Extend horizontal (or vertical) line segments, called whiskers, from each end of the box to the smallest and largest observations in the data set.

Example 4.11 Revisiting Improvement Score Data

Let's reconsider the data from Example 4.9 on improvement score for people in the variable priority practice group. The ordered observations are

Ordered Data

Lower Half:
1200 1300 2300 3200 3300 3800 4000 4100 4300 4800

Median = 5150

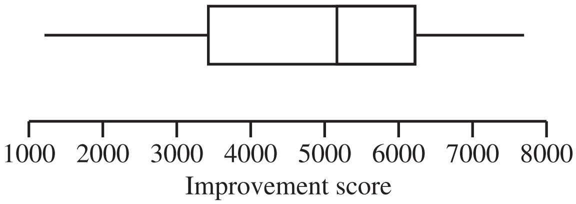
Upper Half:
5500 5700 5700 5800 6000 6300 6800 6800 6900 7700

To construct a boxplot of these data, you need the following information: the smallest observation, the lower quartile, the median, the upper quartile, and the largest observation. This collection of summary measures is often referred to as a **five-number summary**. For this data set

smallest observation = 1200
lower quartile = median of the lower half = 3550
median = average of the 10th and 11th observations in the ordered list = 5150
upper quartile = median of the upper half = 6150
largest observation = 7700

Figure 4.9 shows the corresponding boxplot. The median line is somewhat closer to the upper edge of the box than to the lower edge, suggesting a concentration of values in the upper part of the middle half. The lower whisker is longer than the upper whisker, suggesting that the distribution of improvement scores for the variable practice strategy group is not symmetric.

Figure 4.9
Skeletal boxplot for the improvement score data of Example 4.10.



What Do You Think?

Suppose that a box plot has a much longer upper whisker than the lower whisker. What would that tell you about the distribution of values in the data set?

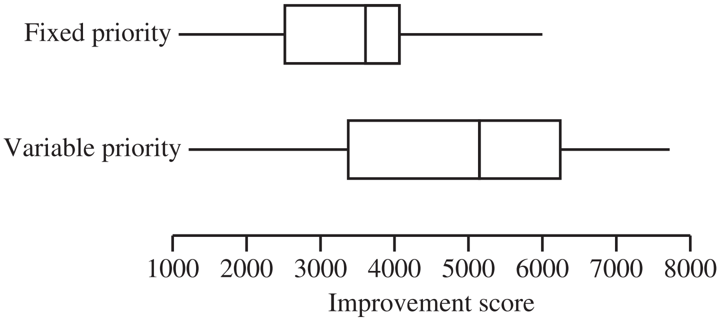
Boxplots are often used to compare groups. For example, you could compare video game improvement scores for the two different practice strategies described in Example 4.9 using a **comparative boxplot**.

Definition

Comparative boxplot: Two or more boxplots drawn using the same numerical scale.

The video game improvement scores for the two different strategy groups were used to construct the comparative boxplot shown in Figure 4.10.

Figure 4.10
Comparative boxplots for improvement scores of fixed priority practice group and variable priority practice group.



From the comparative boxplot, you can see that both data distributions are approximately symmetric. The improvement scores tend to be higher for the variable priority practice group than for the fixed priority practice group. However, there is more consistency in the improvement scores for the fixed priority practice group. Improvement scores in the variable priority practice group are more spread out, indicating more variability in improvement scores for this group.

The sequence of steps used to construct a skeletal boxplot is easily modified to include information about outliers.

Definition

Outlier: An observation that is more than 1.5(iqr) away from the nearest quartile (the nearest end of the box in a boxplot).

An outlier is an **extreme outlier** if it is more than 3(iqr) from the nearest quartile and it is a **mild outlier** otherwise.

Copyright 2025 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s). Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

A **modified boxplot** represents mild outliers by solid circles and extreme outliers by open circles, and the whiskers extend on each end to the most extreme observations that are *not* outliers.

Construction of a Modified Boxplot

1. Draw a horizontal (or vertical) measurement scale.
2. Construct a rectangular box with a left (or lower) edge at the lower quartile and right (or upper) edge at the upper quartile. This makes the box width equal to the iqr.
3. Draw a vertical (or horizontal) line segment inside the box at the location of the median.
4. Determine if there are any mild or extreme outliers in the data set.
5. Draw whiskers that extend from each end of the box to the most extreme observation that is *not* an outlier.
6. Draw a solid circle to mark the location of any mild outliers in the data set.
7. Draw an open circle to mark the location of any extreme outliers in the data set.

Example 4.12 Golden Rectangles

Understand the context ➤

● The accompanying data came from an anthropological study of rectangular shapes (***Lowie's Selected Papers in Anthropology,*** **Cora Dubios, ed. [Berkeley, CA: University of California Press, 1960]: 137–142)**. Observations were made on the variable x = width/length for a sample of $n = 20$ beaded rectangles used in Shoshoni Indian leather handicrafts:

Consider the data ➤

0.553	0.570	0.576	0.601	0.606	0.606	0.609	0.611	0.615	0.628
0.654	0.662	0.668	0.670	0.672	0.690	0.693	0.749	0.844	0.933

The quantities needed for constructing the modified boxplot have been calculated and are shown here:

Do the work ➤

median = 0.641
lower quartile = 0.606
upper quartile = 0.681

iqr = 0.681 − 0.606 = 0.075
1.5(iqr) = 0.1125
3(iqr) = 0.225

Then,

$$\text{(upper quartile)} + 1.5(\text{iqr}) = 0.681 + 0.1125 = 0.7935$$
$$\text{(lower quartile)} - 1.5(\text{iqr}) = 0.606 - 0.1125 = 0.4935$$

This means that 0.844 and 0.933 are both outliers on the upper end (because they are greater than 0.7935). There are no outliers on the lower end (because no observations are less than 0.4935). Because

$$\text{(upper quartile)} + 3(\text{iqr}) = 0.681 + 0.225 = 0.906$$

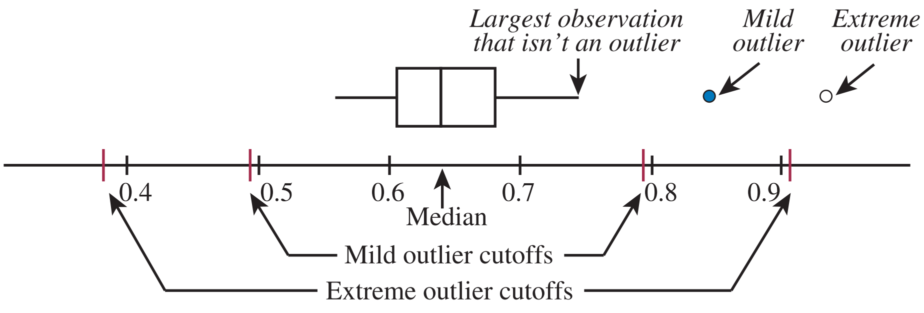
Interpret the results ➤

0.933 is an extreme outlier and 0.844 is only a mild outlier. The upper whisker extends to the largest observation that is not an outlier, which is 0.749. The lower whisker extends to the smallest observation that is not an outlier, which is 0.553. The boxplot is shown in Figure 4.11. The median line is not at the center of the box, so there is a slight asymmetry in the middle half of the data. However, the most striking feature is the presence of the two outliers. These two x values considerably exceed the "golden ratio" of 0.618, used since antiquity as an aesthetic standard for rectangles.

● **Data set available online**

Copyright 2025 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s). Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

Figure 4.11
Boxplot for the rectangle data in Example 4.11.



Example 4.13 College Degrees

Understand the context ➤

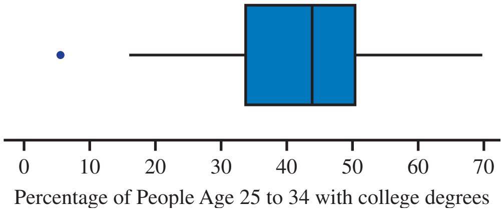
● The accompanying data on the percentage of 25- to 34-year-old people who have completed a 4-year college degree for 43 countries in 2019 are from the **Organisation for Economic Co-operation and Development (data.oecd.org)**. The data values have been arranged in order from smallest (South Africa at 5.6%) to greatest (Korea at 69.8%). The percentage for the United States is 50.4.

Consider the data ➤

5.6	16.1	21.3	21.8	25.6	30.0	30.6	31.4	32.6	33.3	**33.7**
35.3	37.4	39.2	40.0	41.6	41.8	42.4	42.8	43.5	43.8	**43.8**
44.1	46.5	47.0	47.1	47.1	47.3	48.1	48.4	48.7	49.1	**50.4**
51.8	52.5	52.7	55.2	55.5	56.0	60.1	62.1	63.0	69.8	

Figure 4.12 shows a Minitab boxplot for the college degree data. Note that the lengths of the two whiskers are about the same, but that there is an outlier on the low end (South Africa with 5.6% of people in this age group with college degrees).

Figure 4.12
Minitab boxplot for the degree data of Example 4.13.



Notice that Minitab does not distinguish between mild outliers and extreme outliers in the boxplot. For the college degree data,

Do the work ➤

lower quartile = 33.7
upper quartile = 50.4
iqr = 50.4 − 33.7 = 16.7

Then

1.5(iqr) = 25.05
3(iqr) = 50.10

You can calculate outlier boundaries as follows:

lower quartile − 1.5(iqr) = 33.7 − 25.05 = 8.65
lower quartile − 3(iqr) = 33.7 − 50.10 = −16.40

The observation for South Africa (5.6) is a mild outlier because it is less than 8.65 (the lower quartile − 1.5(iqr)) but not less than −16.40 (the lower quartile − 3(iqr)). There are no extreme outliers in this data set.

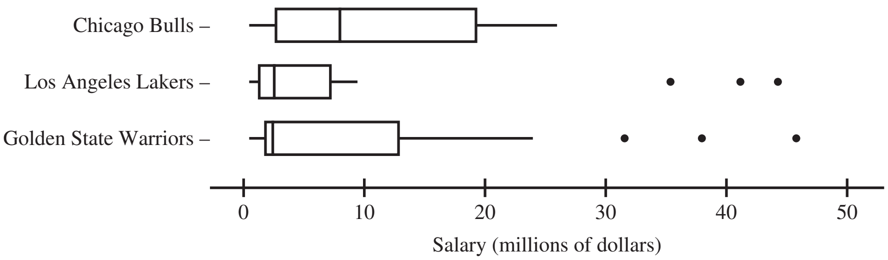
● **Data set available online**

Example 4.14 NBA Player Salaries

Understand the context ➤

The website **HoopsHype (hoopshype.com/salaries)** publishes salaries of NBA players. The 2020–2021 salaries of players on three teams were used to construct the comparative boxplot shown in Figure 4.13.

Figure 4.13
Comparative boxplot for salaries of three NBA teams



Interpret the results ➤

The comparative boxplot reveals some interesting similarities and differences in the salary distributions of the three teams. The minimum salary was actually the same for all three teams. The median salary was greatest for the Bulls, but the other two teams both had players with salaries that were much greater than the highest Bulls' salary. The Lakers and the Warriors each have three players with salaries that are much higher than the other players on the team, whereas there are no outliers in the salaries for the Bulls. With the exception of the three outliers, the Lakers salaries show less variability than the salaries of the other two teams, which is reflected in the smaller interquartile range (the width of the box). The salary distribution for the Bulls looks much more symmetric than the distribution of the other two teams and the salary distribution for the Warriors is quite skewed.

Exercises 4.34–4.41

● **Data set available online**

4.34 Based on a large national sample of working adults, the **U.S. Census Bureau** reports the following information on travel time to work for those who do not work at home:

lower quartile = 7 minutes
median = 18 minutes
upper quartile = 31 minutes

Also given was the mean travel time, which was reported as 22.4 minutes.

a. Is the travel time distribution more likely to be approximately symmetric, positively skewed, or negatively skewed? Explain your reasoning based on the given summary quantities.

b. Suppose that the minimum travel time was 1 minute and that the maximum travel time in the sample was 205 minutes. Construct a skeletal boxplot for the travel time data. (Hint: See Example 4.11.)

c. Were there any mild or extreme outliers in the data set? How can you tell? (Hint: See Example 4.12.)

4.35 ● Internet users rated their Internet service providers on value, reliability, and speed. These ratings were combined to get an overall satisfaction score. The scores for 26 different providers (arranged in order from high to low) are given below (***Consumer Reports*, August 2021**).

85 82 71 70 70 70 70 68 67 67 64 64 63
63 63 62 62 61 61 60 58 58 57 56 51 46

a. Calculate the quartiles and the interquartile range.

b. Are there outliers in this data set? If so, which observations are outliers?

c. Draw a boxplot for this data set and comment on the interesting features of this plot.

4.36 ● Data on the gasoline tax per gallon (in cents) in 2019 for the 50 U.S. states and the District of Columbia (DC) are shown below **(www.eia.gov)**.

State	Gasoline Tax (cents per gallon)	State	Gasoline Tax (cents per gallon)
Alabama	21.1	Montana	32.3
Alaska	14.4	Nebraska	30.5
Arizona	19.0	Nevada	33.8
Arkansas	21.8	New Hampshire	23.8
California	54.4	New Jersey	41.4
Colorado	22.0	New Mexico	18.9
Connecticut	36.9	New York	44.1
Delaware	23.0	North Carolina	36.5
DC	23.5	North Dakota	23.0
Florida	42.0	Ohio	28.0
Georgia	35.3	Oklahoma	20.0
Hawaii	46.4	Oregon	36.8
Idaho	33.0	Pennsylvania	58.7
Illinois	32.0	Rhode Island	34.0
Indiana	42.9	South Carolina	20.8
Iowa	30.7	South Dakota	30.0
Kansas	24.0	Tennessee	26.4
Kentucky	26.0	Texas	20.0
Louisiana	20.0	Utah	30.0
Maine	30.0	Vermont	31.2
Maryland	35.3	Virginia	20.7
Massachusetts	26.5	Washington	49.4
Michigan	38.4	West Virginia	35.7
Minnesota	28.6	Wisconsin	32.9
Mississippi	18.8	Wyoming	24.0
Missouri	17.4		

a. The smallest value in the data set is 14.4 (Alaska), and the largest value is 58.7 (Pennsylvania). Are these values outliers? Explain. (Hint: See Example 4.12.)

b. Construct a boxplot for this data set and comment on the interesting features of the plot.

4.37 ● The National Center for Health Statistics reported the estimated percentage of households with only wireless phone service (no landline) in 2018 for the 50 states and the District of Columbia (DC) **(cdc.gov)**. In the accompanying data table, each state was also classified into one of three geographical regions—West (W), Middle (M), and East (E).

Wireless %	Region	State	Wireless %	Region	State
57.6	M	AL	51.9	W	MT
57.3	W	AK	58.7	M	NE
64.9	W	AZ	62.7	W	NV
66.4	M	AR	58.4	M	ND
55.4	W	CA	46.0	E	NH
61.4	W	CO	38.1	E	NJ
40.8	E	CT	66.4	W	NM
45.9	E	DE	38.8	E	NY
57.4	E	DC	57.3	E	NC
60.9	E	FL	58.5	E	OH
59.0	E	GA	66.7	M	OK
51.6	W	HI	63.4	W	OR
72.2	W	ID	43.4	E	PA
59.8	M	IL	45.6	E	RI
63.3	M	IN	55.7	E	SC
61.3	M	IA	60.9	M	SD
64.6	M	KS	61.2	M	TN
61.9	M	KY	67.9	M	TX
57.6	M	LA	69.0	W	UT
47.7	E	ME	49.1	E	VA
39.4	E	MD	48.0	E	VT
39.9	E	MA	58.2	W	WA
57.3	M	MI	52.6	E	WV
56.2	M	MN	54.0	M	WI
66.5	M	MS	69.5	W	WY
64.2	M	MO			

a. Construct a comparative boxplot that makes it possible to compare wireless percent for the three geographical regions.

b. Does the graphical display in Part (a) reveal any striking differences, or are the distributions similar for the three regions?

4.38 ● Fiber content (in grams per serving) for 18 high-fiber cereals **(consumerreports.org)** are shown below.

Fiber Content

7	10	10	7	8	7	12	12	8
13	10	8	12	7	14	7	8	8

a. Find the median, quartiles, and interquartile range for the fiber content data set.

b. Explain why the minimum value for the fiber content data set and the lower quartile for the fiber content data set are equal.

4.39 ● In addition to the fiber contents given in the previous exercise, sugar content (in grams per serving) for the same 18 high-fiber cereals were also given **(consumerreports.org)**.

11 6 14 13 0 18 9 10 19
6 10 17 10 10 0 9 5 11

a. Calculate the median, quartiles, and interquartile range for the sugar content data.
b. Are there any outliers in the sugar content data?

4.40 ● The paper **"Gambling with Your Health, Part 2: The Effect of Two Sanitation Procedures on the Disinfection of Bacterial Contamination on Casino Cheques" (*Journal of Environmental Health* [2020]: 8–14)** gave the following data on the number of bacterial colonies detected on casino poker chips after being treated with a disinfectant or run through a dishwasher.

Disinfectant

117 232 9 48 47 55 38 25 65
63 68 98 152 46 78 80 78 55

Dishwasher

22 11 13 6 5 5 34 28 44 10 13
9 5 25 9 2 13 6 13 10 9

Construct a comparative boxplot and comment on the differences and similarities in the two data distributions. Which cleaning method would you recommend and why?

4.41 ● The article **"The Best—and Worst—Places to be a Working Woman" (*The Economist*, Graphic Detail for March 3, 2016)** reported values of what it calls the glass-ceiling index, which is designed to rate countries based on women's chances of equal treatment at work. The index weights factors that include participation of women in higher education, participation in the workforce by women, pay, and parental leave policies. The best possible value for this index is 100. Data for 29 countries are shown in the accompanying table.

Country	Glass-Ceiling Index
Iceland	82.8
Norway	79.3
Sweden	79.0
Finland	73.8
Hungary	70.4
Poland	70.0
France	68.3
Denmark	66.6
New Zealand	63.8
Belgium	63.3
Canada	62.3
Portugal	60.8
Spain	59.1
Israel	58.3
Slovakia	57.8
Austria	57.1
Germany	57.0
Australia	56.2
United States	55.9
Czech Republic	55.1
Italy	53.7
Netherlands	51.8
Greece	50.8
Great Britain	50.5
Ireland	48.4
Switzerland	40.6
Japan	28.8
Turkey	27.2
South Korea	25.0

a. Are there outliers in this data set? If so, which observations are outliers?
b. Draw a modified boxplot for this data set.
c. The article points out that Nordic countries (Iceland, Sweden, Norway, and Finland) come out on top on this index. Where are the values for the Nordic countries located in terms of the boxplot?

Section 4.4 Interpreting Center and Variability: Chebyshev's Rule, the Empirical Rule, and *z* Scores

The mean and standard deviation can be combined to make informative statements about how the values in a data set are distributed and about the relative position of a particular value in a data set. To do this, it is useful to be able to describe how far away a particular observation is from the mean in terms of the standard deviation. For example, you might say that an observation is 2 standard deviations above the mean or that an observation is 1.3 standard deviations below the mean.

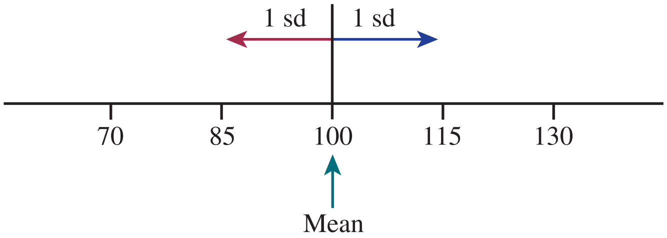
Example 4.15 Standardized Test Scores

Consider a data set of scores on a standardized test with a mean and standard deviation of 100 and 15, respectively. You can make the following statements:

1. Because $100 - 15 = 85$, you would say that a score of 85 is "1 standard deviation *below* the mean." Similarly, $100 + 15 = 115$ is "1 standard deviation *above* the mean" (see Figure 4.14).

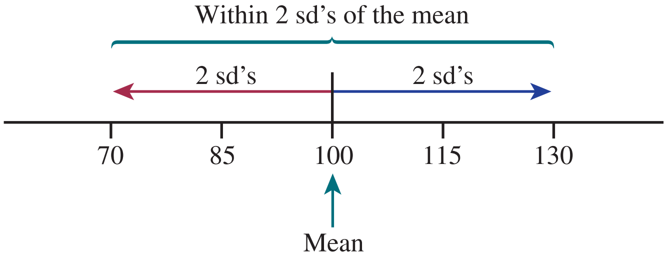
Copyright 2025 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s). Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

Figure 4.14
Values within 1 standard deviation of the mean (Example 4.15).



2. Because 2 times the standard deviation is $2(15) = 30$, scores between 70 and 130 are those *within* 2 standard deviations of the mean (see Figure 4.15).

Figure 4.15
Values within 2 standard deviations of the mean (Example 4.15).



3. Because $100 + (3)(15) = 145$, scores above 145 are greater than the mean by more than 3 standard deviations.

Sometimes in published articles, the mean and standard deviation are reported, but a graphical display of the data is not given. However, using a result called Chebyshev's Rule, it is possible to get a sense of the distribution of data values based only on knowing the mean and the standard deviation.

Chebyshev's Rule

Consider any number k, where $k \geq 1$. The percentage of observations that are within k standard deviations of the mean is *at least* $100\left(1 - \frac{1}{k^2}\right)\%$.

Substituting selected values of k into Chebyshev's Rule gives the following results.

Number of Standard Deviations, k	$1 - \frac{1}{k^2}$	Percentage Within k Standard Deviations of the Mean
2	$1 - \frac{1}{4} = 0.75$	at least 75%
3	$1 - \frac{1}{9} = 0.89$	at least 89%
4	$1 - \frac{1}{16} = 0.94$	at least 94%
4.472	$1 - \frac{1}{20} = 0.95$	at least 95%
5	$1 - \frac{1}{25} = 0.96$	at least 96%
10	$1 - \frac{1}{100} = 0.99$	at least 99%

Copyright 2025 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s). Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

Example 4.16 Child Care Costs

Understand the context ➤

The report **"Child Care Costs by State 2023" (worldpopulationreview.com, 2023)** gives the average annual cost of childcare in 2023 for the 50 U.S. states and the District of Columbia. For this data set, the mean was \$11,422 and the standard deviation was \$3680. A plot of the data showed that the data distribution was skewed to the right, with two noticeable outliers (the District of Columbia and Massachusetts). Figure 4.16 displays values that are 1, 2, and 3 standard deviations from the mean.

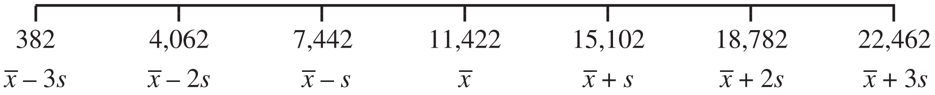


Figure 4.16
Measurement scale for child care cost (Example 4.16).

Chebyshev's Rule allows you to assert the following:

1. At least 75% of the states must have annual child care costs that are between \$4,062 and \$18,782 hours (within 2 standard deviations of the mean).
2. Because at least 89% of the observations must be between \$382 and \$22,462, at most 11% are outside this interval. Since it is unlikely that a state would have an average annual child care cost of less than \$382, you can conclude that at most, 11% of states have an annual average child care cost of more than \$22,462.
3. The values \$5,902 and \$16,942 are 1.5 standard deviations to either side of $\bar{x}$, so using $k = 1.5$ in Chebyshev's Rule implies that at least 55.6% of the observations must be between these two values. This means that at most 44.4% of the observations are less than \$5,902—*not* at most 22.2%, because the distribution of values may not be symmetric.

Because Chebyshev's Rule can be used with any data set (distribution), whether symmetric or skewed, you must be careful when making statements about the proportion above a particular value, below a particular value, or inside or outside an interval that is not centered at the mean. The rule must be used in a conservative fashion. There is another aspect of this conservatism. The rule states that *at least* 75% of the observations are within 2 standard deviations of the mean, but for many data sets substantially more than 75% of the values satisfy this condition. The same sort of understatement is also frequently encountered for other values of k (numbers of standard deviations).

Example 4.17 Penguin Body Mass

Understand the context ➤

Figure 4.17 gives a stem-and-leaf display of flipper length (in mm) of 150 Adelie penguins **("Ecological Sexual Dimorphism and Environmental Variability within a Community of Antarctic Penguins" (journals.plos.org, 2014)**.

Summary quantities include

$\bar{x} = 190 \qquad s = 7 \qquad 2s = 14 \qquad 3s = 21$

Figure 4.17
Stem-and-leaf display of flipper lengths for Example 4.17.

```
17 | 2468889
18 | 0000011111222334444444555555556666666777777777788888999999
19 | 00000000000000000001111111111222223333333333444555555555556666667777788889999
20 | 00122358
21 | 0
```

Stem: Tens
Leaf: Ones

In Figure 4.17, all of the observations that are within two standard deviations of the mean are shown in blue. Table 4.6 shows how Chebyshev's Rule can sometimes considerably understate actual percentages.

Table 4.6 **Summarizing the Distribution of IQ Scores**

k = Number of sd's	$\bar{x} \pm ks$	Chebyshev	Actual
2	176 to 204	at least 75%	97% (145)
2.5	172.5 to 207.5	at least 84%	98% (147)
3	169 to 211	at least 89%	100% (150)

← *the blue leaves in Figure 4.17*

Empirical Rule

The fact that statements based on Chebyshev's Rule are frequently conservative suggests looking for rules that are less conservative and more precise. One useful rule is the **Empirical Rule**, which can be used whenever the distribution of data values is reasonably well described by a normal curve (distributions that are "mound" shaped).

The Empirical Rule

If the histogram of values in a data set can be reasonably well approximated by a normal curve, then

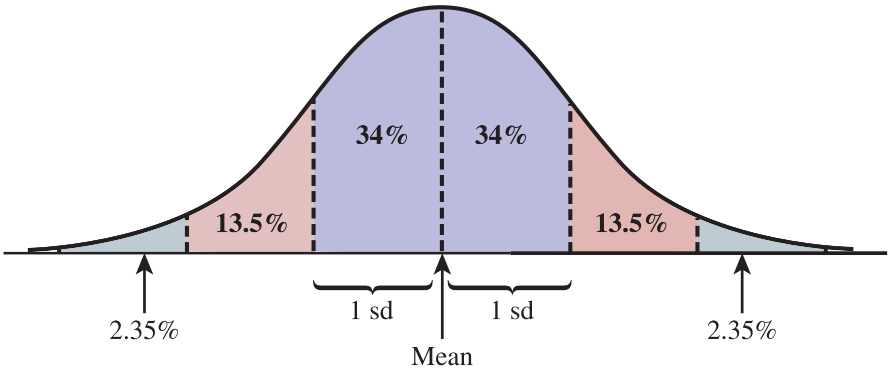
Approximately 68% of the observations are within 1 standard deviation of the mean.

Approximately 95% of the observations are within 2 standard deviations of the mean.

Approximately 99.7% of the observations are within 3 standard deviations of the mean.

The Empirical Rule makes "approximately" instead of "at least" statements, and the percentages for $k = 1, 2,$ and 3 standard deviations are much higher than those of Chebyshev's Rule. Figure 4.18 illustrates the percentages given by the Empirical Rule. In contrast to Chebyshev's Rule, dividing the "within" percentages in half is permissible, because a normal curve is symmetric.

Figure 4.18 Approximate percentages implied by the Empirical Rule.



Example 4.18 Manatee Calf Weights and the Empirical Rule

Understand the context ➤

Manatees are large marine mammals that are sometimes called gentle giants. The authors of the paper **"Prediction Equation of Body Weight of Amazonian Manatee (*Trichechus ignunguis*) Calves in Captivity Using Biometry" (*Natural Science* [2017]: 123–132)** studied manatee calves. Data collected in this study suggest that reasonable values for the mean and standard deviation for body weight of manatee calves are

$$\bar{x} = 32.9 \text{ kg} \qquad s = 4.8 \text{ kg}$$

Graphs in the paper are also consistent with a data distribution that is mound shaped and approximately symmetric. Table 4.7 summarizes what the Empirical Rule tells you about manatee calf body weights.

Table 4.7 **Summarizing the Distribution of Manatee Calf Body Weights**

Number of Standard Deviations	Interval	Percent of Calves with Weights in the Interval
1	32.9 − 1(4.8) to 32.9 + 1(4.8) = 28.1 to 37.7	Approximately 68%
2	32.9 − 2(4.8) to 32.9 + 2(4.8) = 23.3 to 42.5	Approximately 95%
3	32.9 − 3(4.8) to 32.9 + 3(4.8) = 18.5 to 47.3	Approximately 99.7%

Interpret the results ➤ About 68% of manatee calves have body weights between 28.1 and 37.7 kg and it would be very unusual to see a manatee calf with a body weight of less than 18.5 kg or a body weight of more than 47.3 kg.

As you will see in Chapter 7, it is possible to make statements like those of the Empirical Rule for values other than $k = 1, 2,$ or 3 standard deviations. For now, notice that it is unusual to see an observation from a data distribution that is approximately normal that is farther than 2 standard deviations from the mean (only 5%). Also, it is very surprising to see one that is more than 3 standard deviations away.

Measures of Relative Standing

When you obtain your score after taking a test, you probably want to know how it compares to the scores of others who have taken the test. Is your score above or below the mean, and by how much? Does your score place you among the top 5% of those who took the test or only among the top 25%? Questions of this sort are answered by finding ways to measure the position of a particular value in a data set relative to all values in the set. One measure of relative standing is a ***z* score**.

Definition

***z* score:** The ***z* score** corresponding to a particular value is

$$z \text{ score} = \frac{\text{value} - \text{mean}}{\text{standard deviation}}$$

The *z* score tells you how many standard deviations the value is from the mean. It is positive or negative depending on whether the value is greater than or less than the mean.

The process of subtracting the mean and then dividing by the standard deviation is sometimes referred to as standardizing. A *z* score is one example of what is called a standardized score.

Example 4.19 Relatively Speaking, Which Is the Better Offer?

Understand the context ➤ Suppose that two graduating seniors, one a marketing major and one an accounting major, are comparing job offers. The accounting major has an offer for $55,000 per year, and the

Copyright 2025 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s). Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

marketing student has an offer for \$53,000 per year. Summary information about the distribution of offers follows:

Consider the data ➤

Accounting: mean = 56,000 standard deviation = 1500
Marketing: mean = 52,500 standard deviation = 1000

Do the work ➤ Then,

$$\text{accounting } z \text{ score} = \frac{55{,}000 - 56{,}000}{1500} = -0.67$$

(so \$55,000 is 0.67 standard deviation below the mean), but

$$\text{marketing } z \text{ score} = \frac{53{,}000 - 52{,}500}{1000} = 0.5$$

Interpret the results ➤ Relative to the appropriate data sets, the marketing offer is actually more attractive than the accounting offer.

The z score is particularly useful when the distribution of observations is approximately normal. In this case, using the Empirical Rule, a z score outside the interval from -2 to $+2$ occurs for only about 5% of all observations. A z score outside the interval from -3 to $+3$ occurs only about 0.3% of the time.

Percentiles

A particular observation can be located even more precisely by giving the percentage of the data that fall at or below that observation. For example, suppose 95% of all test scores are at or below 650 and only 5% are above 650. Then 650 is the *95th percentile* of the data set (or of the distribution of scores). Similarly, if 10% of all scores are at or below 400 and 90% are above 400, the value 400 is the *10th percentile*.

Definition

Percentile: For any particular number r between 0 and 100, the **rth percentile** is a value such that r percent of the observations in the data set fall at or below that value.

Figure 4.19 illustrates the 90th percentile.

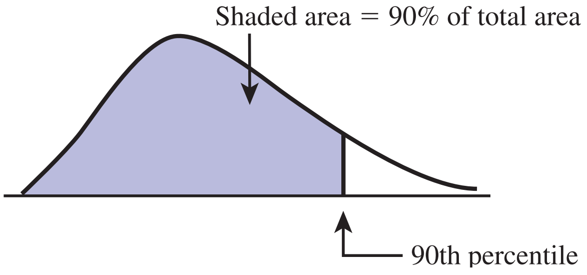


Figure 4.19
Ninetieth percentile for a smoothed histogram.

What Do You Think?

The median of a data set corresponds to the 50th percentile. To what percentile does the lower quartile correspond?

The median is the 50th percentile, and the lower and upper quartiles are the 25th and 75th percentiles, respectively.

Example 4.20 Head Circumference at Birth

Understand the context ➤ In addition to weight and length, head circumference is another measure of health in newborn babies. **The National Center for Health Statistics** reports the following summary values for head circumference (in cm) at birth for males (approximate values read from graphs on the Centers for Disease Control and Prevention website **(cdc.gov)**:

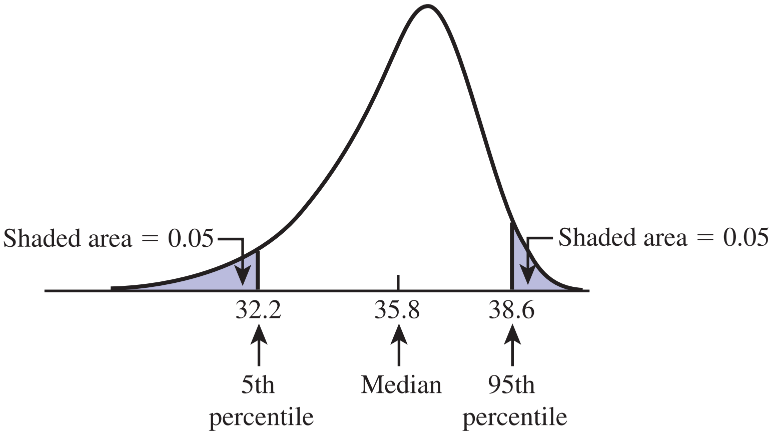
Consider the data ➤

Percentile	5	10	25	50	75	90	95
Head Circumference (cm)	32.2	33.2	34.5	35.8	37.0	38.2	38.6

Interpret the results ➤ Interpreting these percentiles, you know that half of male newborns have head circumferences of less than 35.8 cm, because 35.8 is the 50th percentile (the median). The middle 50% of male newborns have head circumferences between 34.5 cm and 37.0 cm, with about 25% of the head circumferences less than 34.5 cm and about 25% greater than 37.0 cm.

You can tell that the head circumference distribution for male newborns is not symmetric, because the 5th percentile is 3.6 cm below the median, whereas the 95th percentile is only 2.8 cm above the median. This suggests that the bottom part of the distribution stretches out more than the top part of the distribution. This would be consistent with a distribution that is negatively skewed, as shown in Figure 4.20.

Figure 4.20
Negatively skewed distribution.



Exercises 4.42–4.58

● **Data set available online**

4.42 The average playing time of music albums in a large collection is 35 minutes, and the standard deviation is 5 minutes.

a. What value is 1 standard deviation above the mean? 1 standard deviation below the mean? What values are 2 standard deviations away from the mean? (Hint: See Example 4.15.)

b. Without assuming anything about the distribution of times, at least what percentage of the times are between 25 and 45 minutes? (Hint: See Example 4.16.)

c. Without assuming anything about the distribution of times, what can be said about the percentage of times that are either less than 20 minutes or greater than 50 minutes?

d. Assuming that the distribution of times is approximately normal, about what percentage of times are between 25 and 45 minutes? less than 20 minutes or greater than 50 minutes? less than 20 minutes?

4.43 In a study investigating the effect of car speed on accident severity, 5000 reports of fatal automobile accidents were examined, and the vehicle speed at impact was recorded for each one. For these 5000 accidents, the average speed was 42 mph and the standard deviation was 15 mph. A histogram revealed that the vehicle speed at impact distribution was approximately normal.

a. Approximately what proportion of these vehicle speeds were between 27 and 57 mph? (Hint: See Example 4.18.)

Copyright 2025 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s). Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

b. Approximately what proportion of these vehicle speeds exceeded 57 mph?

4.44 **The U.S. Census Bureau (census.gov)** reported the following relative frequency distribution for travel time to work in 2019 for a large sample of adults who did not work at home:

Travel Time (minutes)	Relative Frequency
0 to < 5	0.03
5 to < 10	0.09
10 to < 15	0.13
15 to < 20	0.15
20 to < 25	0.14
25 to < 30	0.07
30 to < 35	0.14
35 to < 40	0.03
40 to < 45	0.04
45 to < 60	0.08
60 to < 90	0.07
90 or more	0.03

a. Draw the histogram for the travel time distribution. In constructing the histogram, assume that the last interval in the relative frequency distribution (90 or more) ends at 200; so the last interval is 90 to < 200. Be sure to use the density scale to determine the heights of the bars in the histogram because not all the intervals have the same width. (Hint: Histograms were covered in Chapter 3.)

b. Describe the interesting features of the histogram from Part (a), including center, shape, and variability.

c. Based on the histogram from Part (a), would it be appropriate to use the Empirical Rule to make statements about the travel time distribution? Explain why or why not.

4.45 For the travel time distribution given in the previous exercise, the approximate mean and standard deviation for the travel time distribution are 32 minutes and 15 minutes, respectively. Based on this mean and standard deviation and the fact that travel time cannot be negative, explain why the travel time distribution could not be well approximated by a normal curve.

4.46 Use the information given in the previous two exercises and Chebyshev's Rule to complete this exercise.

a. Make a statement about

- **i.** the percentage of travel times that were between 0 and 75 minutes
- **ii.** the percentage of travel times that were between 0 and 54 minutes

b. How well do the statements in Part (a) based on Chebyshev's Rule agree with the actual percentages for the travel time distribution? (Hint: You can estimate the actual percentages from the relative frequency distribution given in Exercise 4.44.)

4.47 Mobile homes are tightly constructed for energy conservation. This can lead to a buildup of indoor pollutants. The paper **"A Survey of Nitrogen Dioxide Levels Inside Mobile Homes" (*Journal of the Air Pollution Control Association* [1988]: 647–651)** discussed various aspects of NO_2 concentration in these structures.

a. For one sample of mobile homes in the Los Angeles area, the mean NO_2 concentration in kitchens during the summer was 36.92 ppb, and the standard deviation was 11.34. Making no assumptions about the shape of the NO_2 distribution, what can be said about the percentage of observations between 14.24 and 59.60?

b. Inside what interval can you be sure that at least 89% of the concentration observations will lie?

c. For a sample of mobile homes that were not in Los Angeles, the average kitchen NO_2 concentration during the winter was 24.76 ppb, and the standard deviation was 17.20. Do these values suggest that the histogram of sample observations did not closely resemble a normal curve? (Hint: What is $\bar{x} - 2s$?)

4.48 The article **"Impact of Berkeley Excise Tax on Sugar-Sweetened Beverage Consumption" (*American Journal of Public Health* [2016]: 1865–1871)** estimated that the mean number of times that adults in Berkeley, California, drank regular soda per day to be 0.34. The standard deviation for the number of times per day was estimated to be 0.86. Would you use the Empirical Rule to approximate the proportion of adults who drink regular soda more than 1.20 times per day on average (i.e., the proportion of adults in Berkeley whose value exceeds the mean by more than 1 standard deviation)? Explain your reasoning.

4.49 A student took two national aptitude tests. The national mean and standard deviation were 475 and 100, respectively, for the first test and 30 and 8, respectively, for the second test. The student scored 625 on the first test and 45 on the second test. Use z scores to determine on which exam the student performed better relative to the other test takers. (Hint: See Example 4.19.)

4.50 Suppose that a friend is applying for entrance to college and has taken the SAT. They scored at the 83rd percentile on the verbal section of the test and at the 94th percentile on the math section of the test. Because you have been studying statistics, they ask you for an interpretation of these values. What would you tell them? (Hint: See Example 4.20.)

4.51 **The Center for Microeconomic Data (www.newyorkfed.org/microeconomics)** reported the following percentiles for the amount of student debt for those who had student loans in 2021:

20th percentile = \$5,000
36th percentile = \$10,000
63rd percentile = \$25,000
82nd percentile = \$50,000
90th percentile = \$75,000

For each of these percentiles, write a sentence interpreting the value of the percentile.

4.52 The paper **"Study of the Flying Ability of *Rhynchophorus ferrugineus* Adults Using a Computer-Monitored Mill" (*Bulletin of Entomological Research* [2014]: 462–467)** summarized data from a study of red palm weevils, a pest that is a threat to palm trees. The following frequency distribution from the paper was constructed using the longest flight (in meters) observed for 132 weevils.

Longest Flight (meters)	Frequency (Number of Weevils)
0 to < 100	71
100 to < 2,000	32
2,000 to < 5,000	18
5,000 to < 10,000	8
10,000 or more	3

Estimate the approximate values of the following percentiles:

a. 54th
b. 80th
c. 92nd

4.53 Suppose that the manufacturer of a scale claims that its scale weighs items up to 110 pounds and provides accuracy to within 0.25 ounce. Suppose that a 50-ounce weight was repeatedly weighed on this scale and the weight readings recorded. The mean value was 49.5 ounces, and the standard deviation was 0.1. What can be said about the percentage of the time that the scale actually showed a weight that was within 0.25 ounce of the true value of 50 ounces? (Hint: Use Chebyshev's Rule.)

4.54 Suppose that your statistics professor returned your first midterm exam with only a z score written on it. They also told you that a histogram of the scores was approximately normal. How would you interpret each of the following z scores?

a. 2.2
b. 0.4
c. 1.8
d. 1.0
e. 0

4.55 The paper **"Answer Changing in Multiple Choice Assessment: Change that Answer When in Doubt—and Spread the Word" (*BMC Medical Education* [2007]: 28–32)** reported that for a group of 72 students, the average number of responses changed from the correct answer to an incorrect answer on a test containing 78 multiple-choice items was 0.9. The corresponding standard deviation was reported to be 1.0. Based on this mean and standard deviation, what can you tell about the shape of the distribution of the variable *number of answers changed from right to wrong*? What can you say about the number of students who changed at least three answers from correct to incorrect?

4.56 Suppose that the average reading speed of students completing a speed-reading course is 450 words per minute (wpm). If the standard deviation is 70 wpm, find the z score associated with each of the following reading speeds.

a. 320 wpm
b. 475 wpm
c. 420 wpm
d. 610 wpm

4.57 ● The following data values are 2018 per capita operating expenditures on public libraries for each of the 50 U.S. states and the District of Columbia **(amacad.org)**:

16.42	28.78	36.04	42.42	54.34
18.85	29.58	37.33	45.05	55.62
20.13	29.58	38.11	45.67	57.92
20.30	29.68	38.12	47.28	58.41
21.52	30.42	40.30	47.85	59.92
21.95	31.58	41.20	48.98	61.37
22.67	31.88	41.34	49.34	63.46
26.78	32.83	41.84	49.97	66.09
26.95	34.44	42.14	53.30	66.46
28.08	35.76	42.25	53.91	69.34
84.84				

a. Summarize this data set with a frequency distribution. Construct the corresponding histogram.
b. Use the histogram in Part (a) to find approximate values of the following percentiles:
  i. 50th
  ii. 70th
  iii. 10th
  iv. 90th
  v. 40th

4.58 The accompanying table gives the mean and standard deviation of reaction times (in seconds) for each of two different stimuli:

	Stimulus 1	Stimulus 2
Mean	6.0	3.6
Standard deviation	1.2	0.8

Copyright 2025 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s). Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

If your reaction time is 4.2 seconds for the first stimulus and 1.8 seconds for the second stimulus, to which stimulus are you reacting relatively more quickly (compared with other individuals)? (Hint: See Example 4.19.)

Section 4.5 Interpreting and Communicating the Results of Statistical Analyses

As was the case with the graphical displays of Chapter 3, the summary statistics introduced in this chapter help you better understand the variables under study. If you have collected data on the amount of money students spend on textbooks at a particular university, most likely you did so because you wanted to learn about the distribution of this variable (amount spent on textbooks) for the population of interest (in this case, students at the university). Numerical measures of center and variability and boxplots provide information, and they also allow you to communicate to others what you have learned from the data.

Communicating the Results of Statistical Analyses

When reporting the results of a data analysis, it is common to begin with descriptive information about the variables of interest. It is always a good idea to start with a graphical display of the data, and, as you saw in Chapter 3, graphical displays of numerical data are usually described in terms of center, variability, and shape. The numerical measures of this chapter can help you to be more specific when describing the center and variability of a data set.

When describing center and variability, you must first decide which measures to use. Common choices are to use either the sample mean and standard deviation or the sample median and interquartile range (and maybe even a boxplot). Because the mean and standard deviation can be sensitive to extreme values in the data set, they are best used when the distribution shape is approximately symmetric and when there are few outliers. If the data set is noticeably skewed or if there are outliers, the median and iqr are generally used to describe center and variability.

Interpreting the Results of Statistical Analyses

It is relatively rare to find raw data in published sources. Typically, only a few numerical summary quantities are reported. You need to be able to interpret these values and understand what they tell you about the underlying data set.

For example, a university conducted an investigation of the amount of time required to process an application for admission once an applicant had entered their information online. One of the individuals who performs this task was asked to record starting time and completion time for 50 randomly selected applications. The resulting times (in minutes) were summarized using the mean, median, and standard deviation:

$$\begin{aligned} \bar{x} &= 27.8 \\ \text{median} &= 27.4 \\ s &= 2.14 \end{aligned}$$

What do these summary values tell you about processing times? The average time required was 27.8 minutes. The value of the standard deviation suggests that there was variability in the times and that there were some processing times that differed quite a bit from the mean time. The median tells you that half of the applications required less than 27.4 minutes. The fact that the mean exceeds the median suggests that some unusually large values in the data set may have affected the value of the mean. This last conjecture is confirmed by the stem-and-leaf display of the data given in Figure 4.21.

Figure 4.21
Stem-and-leaf display of application processing times.

24	8
25	02345679
26	0001234566779
27	223556688
28	23334
29	002
30	011168
31	134
32	2
33	
34	3

Stem: Ones
Leaf: Tenths

The administrators conducting the study looked at the outlier of 34.3 minutes and at the other relatively large values in the data set. They found that the five largest values came from applications that were entered before lunch. After talking with the individual who entered the data, the administrators speculated that morning entry times might differ from afternoon entry times because there tended to be more distractions and interruptions (phone calls, etc.) during the morning hours, when the admissions office tended to be busier.

When morning and afternoon entry times were separated, the following summary statistics resulted:

Morning (based on $n = 20$ applications):	$\bar{x} = 29.1$	median = 28.7	$s = 2.329$
Afternoon (based on $n = 30$ applications):	$\bar{x} = 27.0$	median = 26.7	$s = 1.529$

Clearly, the mean processing time is higher for applications processed in the morning. The individual processing times also differ more from one another in the mornings than in the afternoons (because the standard deviation for morning processing times, 2.329, is about 1.5 times as large as 1.529, the standard deviation for afternoon processing times).

What to Look for in Published Data

Here are a few questions to ask yourself when you interpret numerical summary measures.

- Is the chosen summary measure appropriate for the type of data collected? In particular, watch for inappropriate use of the mean and standard deviation with categorical data that has simply been coded numerically.
- If both the mean and the median are reported, how do the two values compare? What does this suggest about the distribution of values in the data set? If only the mean or the median was used, was an appropriate measure selected?
- Is the standard deviation large or small? Is the value consistent with your expectations regarding variability? What does the value of the standard deviation tell you about the variable being summarized?
- Can anything of interest be said about the values in the data set by applying Chebyshev's Rule or the Empirical Rule?

For example, consider a study that investigated whether people tend to spend more money when they are paying with a credit card than when they are paying with cash. The authors of the paper **"Monopoly Money: The Effect of Payment Coupling and Form on Spending Behavior" (*Journal of Experimental Psychology: Applied* [2008]: 213–225)** randomly assigned each of 114 volunteers to one of two experimental groups. Participants were given a menu for a new restaurant that showed nine menu items. They were then asked to estimate the amount they would be willing to pay for each item. A price score was computed for each participant by averaging the nine prices assigned.

The difference between the two experimental groups was that the menu viewed by one group showed a credit card logo at the bottom of the menu while there was no credit card logo on the menu that those in the other group viewed. The following passage appeared in the results section of the paper:

> On average, participants were willing to pay more when the credit card logo was present (M = \$4.53, SD = 1.15) than when it was absent (M = \$4.11, SD = 1.06). Thus, even though consumers were not explicitly informed which payment mode they would be using, the mere presence of a credit card logo increased the price that they were willing to pay.

The price score data distribution was also described as mound shaped with no outliers for each of the two groups. Because price score (the average of the prices that a participant assigned to the nine menu items) is a numerical variable, the mean and standard deviation are reasonable measures for summarizing center and variability in the data set.

Although the mean for the credit-card-logo group is higher than the mean for the no-logo group, the two standard deviations are similar, indicating similar variability in price score from person to person for the two groups.

Because the distribution of price score values was mound shaped for each of the two groups, you can use the Empirical Rule to tell you a bit more about the distribution. For example, for those in the group who viewed the menu with a credit card logo, approximately 95% of the price score values would have been between

$$4.53 - 2(1.15) = 4.53 - 2.3 = 2.23$$

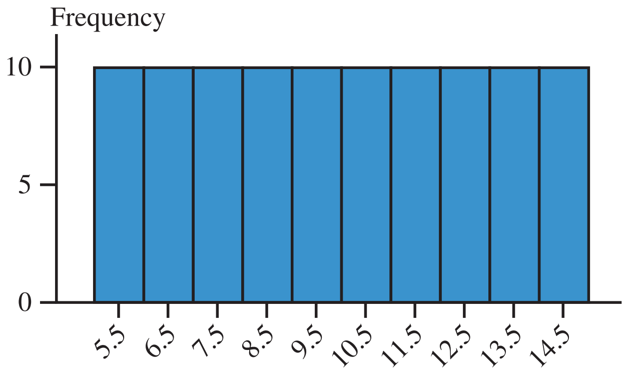
and

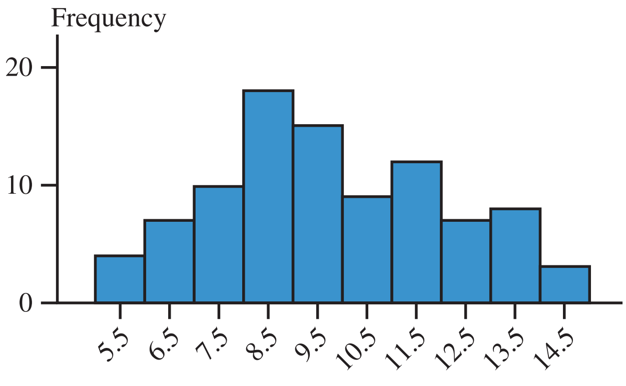
$$4.53 + 2(1.15) = 4.53 + 2.30 = 6.83.$$

A Word to the Wise: Cautions and Limitations

When calculating or interpreting numerical descriptive measures, you need to keep in mind the following:

1. Measures of center don't tell all. Although measures of center, such as the mean and the median, do give a sense of what might be considered a typical value for a variable, this is only one characteristic of a data set. Without additional information about variability and distribution shape, you don't really know much about the behavior of the variable.
2. Data distributions with different shapes can have the same mean and standard deviation. For example, consider the following two histograms:



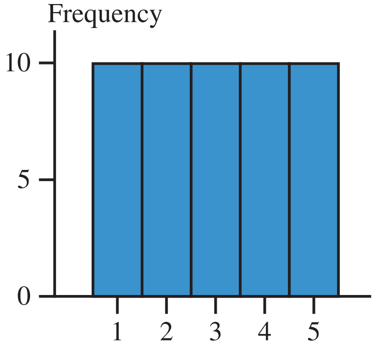


Both histograms summarize data sets that have a mean of 10 and a standard deviation of 2, yet they have very different shapes.

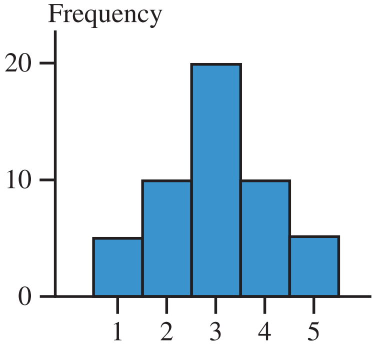
3. Both the mean and the standard deviation are sensitive to extreme values in a data set, especially if the sample size is small. If a data distribution is skewed or if the data set has outliers, the median and the interquartile range may be a better choice for describing center and variability.

4. Measures of center and variability describe the values of the variable studied, not the frequencies in a frequency distribution or the heights of the bars in a histogram. For example, consider the following two frequency distributions and histograms:

Frequency Distribution A		**Frequency Distribution B**	
Value	**Frequency**	**Value**	**Frequency**
1	10	1	5
2	10	2	10
3	10	3	20
4	10	4	10
5	10	5	5



Histogram A



Histogram B

   There is more variability in the data summarized by Frequency Distribution and Histogram A than in the data summarized by Frequency Distribution and Histogram B. This is because the values of the variable described by Histogram and Frequency Distribution B are more concentrated near the mean than are the values for the variable described by Histogram and Frequency Distribution A. Don't be misled by the fact that there is no variability in the frequencies in Frequency Distribution A or the heights of the bars in Histogram A.

5. Be careful with boxplots based on small sample sizes. Boxplots convey information about center, variability, and shape, but when the sample size is small, you should be hesitant to overinterpret shape information. It is really not possible to decide whether a data distribution is symmetric or skewed if only a small sample of observations from the distribution is available.

6. Not all distributions are normal (or even approximately normal). Be cautious in applying the Empirical Rule in situations in which you are not convinced that the data distribution is at least approximately normal. Using the Empirical Rule in such situations can lead to incorrect statements.

7. Watch out for outliers! Unusual observations in a data set often provide important information about the variable under study, so it is important to consider outliers in addition to describing what is typical. Outliers can also be problematic—both because the values of some descriptive measures are influenced by outliers and because some methods for drawing conclusions from data may not be appropriate if the data set has outliers.

Exercises 4.59–4.60

● **Data set available online**

4.59 The authors of the paper **"Delayed Time to Defibrillation after In-Hospital Cardiac Arrest" (*New England Journal of Medicine* [2008]: 9–16)** described a study of how survival is related to the length of time it takes from the time of a heart attack to the administration of defibrillation therapy. The following is a statement from the paper:

> We identified 6789 patients from 369 hospitals who had in-hospital cardiac arrest due to ventricular fibrillation (69.7%) or pulseless ventricular tachycardia (30.3%). Overall, the median time to defibrillation was 1 minute (interquartile range [was] 3 minutes).

Data from the paper on time to defibrillation (in minutes) for these 6789 patients was used to produce the Minitab output and boxplot at the bottom of the page.

a. Why is there no lower whisker in the given boxplot?

b. How is it possible for the median, the lower quartile, and the minimum value in the data set to all be equal? (Note—this is why you do not see a median line in the box part of the boxplot.)

c. The authors of the paper considered a time to defibrillation of greater than 2 minutes as unacceptable. Based on the given boxplot and summary statistics, is it possible that the percentage of patients having an unacceptable time to defibrillation is greater than 50%? Greater than 25%? Less than 25%? Explain.

d. Is the outlier shown at 7 a mild outlier or an extreme outlier?

4.60 The paper **"Portable Social Groups: Willingness to Communicate, Interpersonal Communication Gratifications, and Cell Phone Use among Young Adults" (*International Journal of Mobile Communications* [2007]: 139–156)** describes a study of young adult cell phone use patterns.

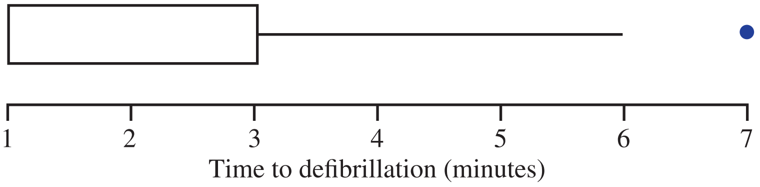
a. Comment on the following quote from the paper. Do you agree with the authors?

> Seven sections of an Introduction to Mass Communication course at a large southern university were surveyed in the spring and fall of 2003. The sample was chosen because it offered an excellent representation of the population under study—young adults.

b. Below is another quote from the paper. In this quote, the author reports the mean number of minutes of cell phone use per week for those who participated in the survey. What additional information would have been provided about cell phone use behavior if the author had also reported the standard deviation?

> Based on respondent estimates, users spent an average of 629 minutes (about 10.5 hours) per week using their cell phone on or off line for any reason.

```
Descriptive Statistics: Time to Defibrillation

Variable     N      Mean     StDev   Minimum       Q1    Median       Q3   Maximum
Time      6789    2.3737    2.0713    1.0000   1.0000    1.0000   3.0000    7.0000
```



Copyright 2025 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s). Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

Chapter Activities

Activity 4.1 Collecting and Summarizing Numerical Data

The purpose of this activity is to collect sample data, summarize the data distribution using both graphical and numerical summaries, and communicate what you have learned about the distribution.

In this activity, you will work in groups to collect data that will provide information about how many hours per week, on average, students at your school spend engaged in a particular activity. You will use the sampling plan designed in Activity 2.1 to collect the data.

1. With your group, pick one of the following activities to be the focus of your study:
   i. Surfing the web
   ii. Studying or doing homework
   iii. Watching TV
   iv. Exercising
   v. Sleeping
   or you may choose a different activity, *subject to the approval of your instructor.*
2. Use the plan developed in Activity 2.1 to collect data on the variable you have chosen for your study.
3. Summarize the resulting data using both numerical and graphical summaries. Be sure to address both center and variability.
4. Write a short article for your school paper summarizing your findings regarding student behavior. Your article should include both numerical and graphical summaries.

Activity 4.2 Airline Passenger Weights

The purpose of this activity is to provide an opportunity to think carefully about the importance of considering variability in a data distribution.

The article **"How Do Airlines Calculate Passengers' Weight?" (www.traveler.com, April 19, 2021)** points out that airlines must estimate the total weight of passengers and carry-on luggage and that the consequences of underestimating the weight of a plane are potentially serious because more weight requires more thrust in order to get off the ground. Rather than weighing passengers, airlines use estimates of average passenger weights.

Although an airplane can fly if it is somewhat overweight if all systems are working properly, if one of the plane's engines fails an overweight plane becomes difficult for the pilot to control.

Assuming that the estimate used for average passenger weight is accurate, discuss the following questions with a partner and then write a paragraph that answers these questions.

1. What role does variability in passenger weights play in creating a potentially dangerous situation for an airline?
2. Would an airline have a lower risk of a potentially dangerous situation if the variability in passenger weight is large or if it is small?

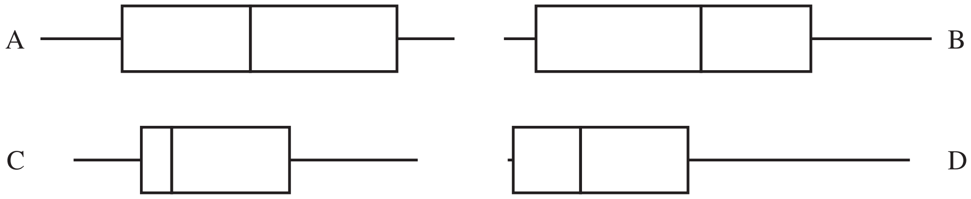
Activity 4.3 Boxplot Shapes

The purpose of this activity is to provide an opportunity to reflect on how boxplots summarize the values in the five-number summary.

In this activity, you will investigate the relationship between boxplot shapes and the corresponding five-number summary. The accompanying figure shows four boxplots, labeled A–D. Also given are 4 five-number summaries, labeled I–IV. Match each five-number summary to the appropriate boxplot. Note that scales are not included on the boxplots, so you will have to think about what the five-number summary implies about characteristics of the boxplot.

Five-Number Summaries

	I	II	III	IV
Minimum	40	4	0.0	10
Lower quartile	45	8	0.1	34
Median	71	16	0.9	44
Upper quartile	88	25	2.2	82
Maximum	106	30	5.1	132



Copyright 2025 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s). Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

Summary Key Concepts and Formulas

Term or Formula	Comment
$x_1, x_2, \ldots, x_n$	Notation for sample data consisting of observations on a variable *x*, where *n* is the sample size.
Sample mean, $\bar{x}$	The most frequently used measure of center of a sample. It can be very sensitive to the presence of even a single outlier (unusually large or small observation).
Population mean, μ	The average *x* value in the entire population.
Sample median	The middle value in the ordered list of sample observations. (For *n* even, the median is the average of the two middle values.) It is very insensitive to outliers.
Deviations from the mean: $x_1 - \bar{x}, x_2 - \bar{x}, \ldots, x_n - \bar{x}$	Quantities used to assess variability in a sample. Except for rounding effects, $\Sigma(x - \bar{x}) = 0$.
The sample variance $s^2 = \dfrac{\Sigma(x - \bar{x})^2}{n - 1}$ **and standard deviation** $s = \sqrt{s^2}$	The most frequently used measures of variability for sample data.
The population variance σ^2 and standard deviation σ	Measures of variability for the entire population.
Quartiles and the interquartile range	The lower quartile separates the smallest 25% of the data from the remaining 75%, and the upper quartile separates the largest 25% from the smallest 75%. The interquartile range (iqr), a measure of variability less sensitive to outliers than *s*, is the difference between the upper and lower quartiles.
Chebyshev's Rule	This rule states that for any number $k \geq 1$, *at least* $100\left(1 - \dfrac{1}{k^2}\right)\%$ of the observations in *any* data set are within *k* standard deviations of the mean. It is typically conservative in that the actual percentages are often considerably greater than the stated lower bound.
Empirical Rule	This rule gives the approximate percentage of observations within 1 standard deviation (68%), 2 standard deviations (95%), and 3 standard deviations (99.7%) of the mean when the data distribution is well approximated by a normal curve.
***z* score**	This quantity gives the distance between an observation and the mean expressed as a certain number of standard deviations. It is positive (negative) if the observation is greater than (less than) the mean.
***r* th percentile**	The value such that *r*% of the observations in the data set fall at or below that value.
Five-number summary	A summary of a data set that includes the minimum, lower quartile, median, upper quartile, and maximum.
Boxplot	A picture that conveys information about the most important features of a numerical data set: center, variability, extent of skewness, and presence of outliers.

Chapter Review Exercises 4.61–4.70

● **Data set available online**

4.61 Acrylamide (a possible cancer-causing substance) forms in high-carbohydrate foods cooked at high temperatures and acrylamide levels can vary widely even within the same type of food. An article appearing in the journal **Food Chemistry (March 2014, pages 204–211)** included the following acrylamide content (in nanograms/gram) for five brands of biscuits:

345 292 334 276 248

a. Calculate the mean acrylamide level and the five deviations from the mean.

b. Verify that, except for the effect of rounding, the sum of the deviations from the mean is equal to 0 for this data set. (If you rounded the sample mean or the deviations, your sum may not be exactly zero, but it should be close to zero if you have calculated the deviations correctly.)

c. Calculate the variance and standard deviation for this data set.

4.62 ● The accompanying data are the prices (in dollars) of 25 tablet computers rated as excellent or good by *Consumer Reports* **(consumerreports.org, September 2022)**.

750	900	1200	1000	300	600
700	630	1280	1300	990	680
350	330	400	270	150	180
230	160	280	210	170	130
750					

a. Use these data to construct a boxplot.

b. Write a few sentences describing the important characteristics of the boxplot.

4.63 Because some homes have selling prices that are much higher than most, the median price is usually used to describe a "typical" home price for a given location. The three accompanying quotes are all from the ***San Luis Obispo Tribune***, but each gives a different interpretation of the median price of a home in San Luis Obispo County. Comment on each of these statements. (Look carefully. At least one of the statements is incorrect.)

a. "So we have gone from 23% to 27% of county residents who can afford the median priced home at $278,380 in SLO County. That means that half of the homes in this county cost less than $278,380 and half cost more." **(October 11, 2001)**

b. "The county's median price rose to $285,170 in the fourth quarter, a 9.6% increase from the same period a year ago, the report said. (The median represents the midpoint of a range.)" **(February 13, 2002)**

c. "'Your median is going to creep up above $300,000 if there is nothing available below $300,000,' Walker said." **(February 26, 2002)**

4.64 ● Although bats are not known for their eyesight, they are able to locate prey (mainly insects) by emitting high-pitched sounds and listening for echoes. A paper appearing in ***Animal Behaviour*** **("The Echolocation of Flying Insects by Bats" [1960]: 141–154)** gave the following distances (in centimeters) at which a bat first detected a nearby insect:

62 23 27 56 52 34 42 40 68 45 83

a. Calculate the sample mean distance at which the bat first detects an insect.

b. Calculate the sample variance and standard deviation for this data set. Interpret these values.

4.65 For the data in the previous exercise, subtract 10 from each sample observation. For the new set of values, calculate the mean and the deviations from the mean. How do these deviations compare to the deviations from the mean for the original sample? How does s^2 for the new values compare to s^2 for the old values? In general, what effect does subtracting (or adding) the same number to each observation have on s^2 and s? Explain.

4.66 For the data of Exercise 4.64, multiply each data value by 10. How does s for the new values compare to s for the original values? More generally, what happens to s if each observation is multiplied by the same positive constant c?

4.67 ● The Bloomberg website included the data in the accompanying table on the number of movies made by 25 *Saturday Night Live* cast members as of 2014 **(bloomberg.com)**. Also given was the top grossing movie made by each and the gross income for that movie adjusted for inflation.

Star	Number of movies	Top-grossing movie	Inflation-adjusted gross of top movie (millions of dollars)
Eddie Murphy	38	Shrek 2	550
Dan Aykroyd	56	Ghostbusters	514
Robert Downey, Jr.	57	The Avengers	646
Ben Stiller	50	Meet the Fockers	352
Bill Murray	46	Ghostbusters	514
Adam Sandler	36	The Waterboy	236
Chris Rock	33	Beverly Hills Cop II	322
Will Ferrell	40	Austin Powers: The Spy Who Shagged Me	293
Mike Myers	16	Shrek 2	550
Joan Cusack	48	Toy Story 3	453
Bill Hader	29	Monsters University	274
Rob Schneider	29	Home Alone 2: Lost in New York	294
Brian Doyle-Murray	32	As Good As It Gets	220
Laurie Metcalf	28	Toy Story 3	453
Kristen Wiig	29	Despicable Me 2	376
Molly Shannon	33	How the Grinch Stole Christmas	359
David Koechner	45	Austin Powers: The Spy Who Shagged Me	293
Jon Lovitz	34	Big	231
Amy Poehler	23	Shrek the Third	370
Chevy Chase	32	National Lampoon's Vacation	147
Billy Crystal	26	Monsters University	274
Harry Shearer	27	The Simpsons Movie	210
Randy Quaid	38	Independence Day	464
David Spade	25	Grown Ups	177
Martin Short	28	Madagascar 3: Europe	224

Copyright 2025 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s). Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

Construct a boxplot for the number of movies data and comment on what the boxplot tells you about the distribution of the number of movies data.

4.68 Refer to the data given in the previous exercise.

c. Are there any outliers in the inflation-adjusted gross movie income data? If so, which data values are outliers?

d. Construct a boxplot for the inflation-adjusted gross movie income data.

e. For the inflation-adjusted gross movie income data, the mean is $351.8 million and the median is $322.0 million. What characteristic of the boxplot explains why the mean is greater than the median for this data set?

4.69 ● The article **"Hair Transplants: A Flagship for Medical Tourism" (laingbusinessnews.com, April 28, 2021)** reported that many adult Americans who get hair transplants travel to another country for the procedure because the cost of a transplant can be lower than in the United States. The accompanying data are the cost of a hair transplant (in dollars, based on 2500 grafts) for 11 different countries.

8,050 13,610 2,676 7,372 5,646 3,650
3,202 5,758 3,350 13,751 17,500

a. Calculate the values of the sample mean and the standard deviation.

b. Calculate the upper quartile, the lower quartile, and the interquartile range.

c. The cost for the United States was $13,610. Is this value an outlier??

d. Construct the boxplot for this data set.

4.70 Suppose that the distribution of scores on an exam can be described by a normal curve with mean 100. The 16th percentile of this distribution is 80.

a. What is the 84th percentile?

b. What is the approximate value of the standard deviation of exam scores?

c. What z score is associated with an exam score of 90?

d. What percentile corresponds to an exam score of 140?

e. Do you think there were many scores below 40? Explain.

Technology Notes

Mean

JMP

1. Input the raw data into a column
2. Click **Analyze** and select **Distribution**
3. Click and drag the column name containing the data from the box under **Select Columns** to the box next to **Y, Columns**
4. Click **OK**

Note: These commands also produce the following statistics: standard deviation, minimum, quartile 1, median, quartile 3, and maximum.

Minitab

1. Input the raw data into C1
2. Select **Stat** and choose **Basic Statistics** then choose **Display Descriptive Statistics...**
3. Double-click C1 to add it to the **Variables** list
4. Click **OK**

Note: These commands also produce the following statistics: standard deviation, minimum, quartile 1, median, quartile 3, and maximum.

SPSS

1. Input the raw data into the first column
2. Select **Analyze** and choose **Descriptive Statistics** then choose **Explore...**
3. Highlight the column name for the variable
4. Click the arrow to move the variable to the **Dependent List** box
5. Click **OK**

Note: These commands also produce the following statistics: median, variance, standard deviation, minimum, maximum, interquartile range, and several plots.

Excel

1. Input the raw data into the first column
2. Click on the **Data** ribbon and select **Data Analysis**

Note: If you do not see **Data Analysis** listed on the Ribbon, see the Technology Notes for Chapter 3 for instructions on installing this add-on.

3. Select **Descriptive Statistics** from the dialog box
4. Click **OK**
5. Click in the box next to **Input Range:** and select the data (if you used and selected column titles, check the box next to **Labels in First Row**)
6. Check the box next to **Summary Statistics**
7. Click **OK**

Note: These commands also produce the following statistics: standard error, median, mode, standard deviation, sample variance, range, minimum, and maximum.

Note: You can also find the mean using the Excel function average.

TI-83/84

1. Input the raw data into **L1** (To access lists press the **STAT** key, highlight the option called **Edit...** then press **ENTER**)
2. Press the **STAT** key
3. Use the arrows to highlight **CALC**
4. Highlight **1-Var Stats** and press **ENTER**
5. Press the **2nd** key and then the **1** key
6. Press **ENTER** Press **ENTER** three times

Note: You may need to scroll to view all of the statistics. This procedure also produces the standard deviation, minimum, Q1, median, Q3, maximum.

TI-Nspire

1. Enter the data into a data list (To access data lists select the spreadsheet option and press **enter**)

Note: Be sure to title the list by selecting the top row of the column and typing a title.

2. Press the **menu** key and select **4:Statistics** then select **1:Stat Calculations** then **1:One-Variable Statistics...**
3. Press **OK**
4. For **X1 List,** select the title for the column containing your data from the drop-down menu
5. Press **OK**

Note: You may need to scroll to view all of the statistics. This procedure also produces the standard deviation, minimum, Q1, median, Q3, maximum.

Median

JMP

1. Input the raw data into a column
2. Click **Analyze** and select **Distribution**
3. Click and drag the column name containing the data from the box under **Select Columns** to the box next to **Y, Columns**
4. Click **OK**

Note: These commands also produce the following statistics: mean, standard deviation, minimum, quartile 1, quartile 3, and maximum.

Minitab

1. Input the raw data into C1
2. Select **Stat** and choose **Basic Statistics** then choose **Display Descriptive Statistics...**
3. Double-click C1 to add it to the **Variables** list
4. Click **OK**

Note: These commands also produce the following statistics: mean, standard deviation, minimum, quartile 1, quartile 3, maximum.

SPSS

1. Input the raw data into the first column
2. Select **Analyze** and choose **Descriptive Statistics** then choose **Explore...**
3. Highlight the column name for the variable
4. Click the arrow to move the variable to the **Dependent List** box
5. Click **OK**

Note: These commands also produce the following statistics: mean, variance, standard deviation, minimum, maximum, interquartile range, and several plots.

Excel

1. Input the raw data into the first column
2. Click on the **Data** ribbon and select **Data Analysis**

Note: If you do not see **Data Analysis** listed on the Ribbon, see the Technology Notes for Chapter 3 for instructions on installing this add-on.

3. Select **Descriptive Statistics** from the dialog box
4. Click **OK**
5. Click in the box next to **Input Range:** and select the data (if you used and selected column titles, check the box next to **Labels in First Row**)
6. Check the box next to **Summary Statistics**
7. Click **OK**

Note: These commands also produce the following statistics: mean, standard error, mode, standard deviation, sample variance, range, minimum, and maximum.

Note: You can also find the median using the Excel function **median**.

TI-83/84

1. Input the raw data into **L1** (To access lists press the **STAT** key, highlight the option called **Edit...** then press **ENTER**)
2. Press the **STAT** key
3. Use the arrows to highlight **CALC**
4. Highlight **1-Var Stats** and press **ENTER**
5. Press the **2nd** key and then the **1** key
6. Press **ENTER**

Note: You may need to scroll to view all of the statistics. This procedure also produces the mean, standard deviation, minimum, Q1, Q3, maximum.

TI-Nspire

1. Enter the data into a data list (To access data lists select the spreadsheet option and press **enter**)

   Note: Be sure to title the list by selecting the top row of the column and typing a title.

2. Press the **menu** key and select **4:Statistics** then select **1:Stat Calculations** then **1:One-Variable Statistics...**
3. Press **OK**
4. For **X1 List**, select the title for the column containing your data from the drop-down menu
5. Press **OK**

Note: You may need to scroll to view all of the statistics. This procedure also produces the mean, standard deviation, minimum, Q1, Q3, maximum.

Variance

JMP

1. Input the raw data into a column
2. Click **Analyze** and select **Distribution**

3. Click and drag the column name containing the data from the box under **Select Columns** to the box next to **Y, Columns**
4. Click **OK**
5. Click the red arrow next to the column name
6. Click **Display Options** then select **Customize Summary Statistics**
7. Check the box next to **Variance**
8. Click **OK**

Minitab

1. Input the raw data into C1
2. Select **Stat** and choose **Basic Statistics** then choose **Display Descriptive Statistics...**
3. Double-click C1 to add it to the **Variables** list
4. Click the **Statistics** button
5. Check the box next to **Variance**
6. Click **OK**
7. Click **OK**

SPSS

1. Input the raw data into the first column
2. Select **Analyze** and choose **Descriptive Statistics** then choose **Explore...**
3. Highlight the column name for the variable
4. Click the arrow to move the variable to the **Dependent List** box
5. Click **OK**

Note: These commands also produce the following statistics: mean, median, standard deviation, minimum, maximum, interquartile range, and several plots.

Excel

1. Input the raw data into the first column
2. Click on the **Data** ribbon and select **Data Analysis**

Note: If you do not see **Data Analysis** listed on the Ribbon, see the Technology Notes for Chapter 3 for instructions on installing this add-on.

3. Select **Descriptive Statistics** from the dialog box
4. Click **OK**
5. Click in the box next to **Input Range:** and select the data (if you used and selected column titles, check the box next to **Labels in First Row**)
6. Check the box next to **Summary Statistics**
7. Click **OK**

Note: These commands also produce the following statistics: mean, standard error, median, mode, standard deviation, range, minimum, and maximum.

Note: You can also find the variance using the Excel function **var**.

TI-83/84

The TI-83/84 does not automatically produce the variance; however, this can be determined by finding the standard deviation and squaring it.

TI-Nspire

The TI-Nspire does not automatically produce the variance; however, this can be determined by finding the standard deviation and squaring it.

Standard Deviation

JMP

1. Input the raw data into a column
2. Click **Analyze** and select **Distribution**
3. Click and drag the column name containing the data from the box under **Select Columns** to the box next to **Y, Columns**
4. Click **OK**

Note: These commands also produce the following statistics: mean, minimum, quartile 1, median, quartile 3, and maximum.

Minitab

1. Input the raw data into C1
2. Select **Stat** and choose **Basic Statistics** then choose **Display Descriptive Statistics...**
3. Double-click C1 to add it to the **Variables** list
4. Click **OK**

Note: These commands also produce the following statistics: standard deviation, minimum, quartile 1, median, quartile 3, maximum.

SPSS

1. Input the raw data into the first column
2. Select **Analyze** and choose **Descriptive Statistics** then choose **Explore...**
3. Highlight the column name for the variable
4. Click the arrow to move the variable to the **Dependent List**box
5. Click **OK**

Note: These commands also produce the following statistics: mean, median, variance, minimum, maximum, interquartile range, and several plots.

Excel

1. Input the raw data into the first column
2. Click on the **Data** ribbon and select **Data Analysis**

Note: If you do not see **Data Analysis** listed on the Ribbon, see the Technology Notes for Chapter 3 for instructions on installing this add-on.

3. Select **Descriptive Statistics** from the dialog box
4. Click **OK**
5. Click in the box next to **Input Range:** and select the data (if you used and selected column titles, check the box next to **Labels in First Row**)
6. Check the box next to **Summary Statistics**
7. Click **OK**

Note: These commands also produce the following statistics: mean, standard error, median, mode, sample variance, range, minimum, and maximum.

Note: You can also find the standard deviation using the Excel function **sd**.

TI-83/84

1. Input the raw data into **L1** (To access lists press the **STAT** key, highlight the option called **Edit...** then press **ENTER**)
2. Press the **STAT** key
3. Use the arrows to highlight **CALC**

4. Highlight **1-Var Stats** and press **ENTER**
5. Press the **2nd** key and then the **1** key
6. Press **ENTER**

Note: You may need to scroll to view all of the statistics. This procedure also produces the mean, minimum, Q1, median, Q3, maximum.

TI-Nspire

1. Enter the data into a data list (To access data lists select the spreadsheet option and press **enter**)

   Note: Be sure to title the list by selecting the top row of the column and typing a title.

2. Press the **menu** key and select **4:Statistics** then select **1:Stat Calculations** then **1:One-Variable Statistics...**
3. Press **OK**
4. For **X1 List,** select the title for the column containing your data from the drop-down menu
5. Press **OK**

Note: You may need to scroll to view all of the statistics. This procedure also produces the mean, minimum, Q1, median, Q3, maximum.

Quartiles

JMP

1. Input the raw data into a column
2. Click **Analyze** and select **Distribution**
3. Click and drag the column name containing the data from the box under **Select Columns** to the box next to **Y, Columns**
4. Click **OK**

Note: These commands also produce the following statistics: mean, standard deviation, minimum, median, and maximum.

Minitab

1. Input the raw data into C1
2. Select **Stat** and choose **Basic Statistics** then choose **Display Descriptive Statistics...**
3. Double-click C1 to add it to the **Variables** list
4. Click **OK**

Note: These commands also produce the following statistics: standard deviation, minimum, quartile 1, median, quartile 3, maximum.

SPSS

1. Input the raw data into the first column
2. Select **Analyze** and choose **Descriptive Statistics** then choose **Frequencies...**
3. Highlight the column name for the variable
4. Click the arrow to move the variable to the **Dependent List** box
5. Click the **Statistics** button
6. Check the box next to Quartiles
7. Click **Continue**
8. Click **OK**

Excel

1. Input the raw data into the first column
2. Select the cell where you would like to place the first quartile results
3. Type "=QUARTILE(array, quart)" where array is the range of cells containing the raw data. If the first quartile is desired, type 1 for quart. Press **ENTER**

Note: To find the third quartile, type 3 for quart in Step 3.

TI-83/84

1. Input the raw data into **L1** (To access lists press the **STAT** key, highlight the option called **Edit...** then press **ENTER**)
2. Press the **STAT** key
3. Use the arrows to highlight **CALC**
4. Highlight **1-Var Stats** and press **ENTER**
5. Press the **2nd** key and then the **1** key
6. Press **ENTER**

Note: You may need to scroll to view all of the statistics. This procedure also produces the mean, standard deviation, minimum, median, maximum.

TI-Nspire

1. Enter the data into a data list (To access data lists select the spreadsheet option and press **enter**)

Note: Be sure to title the list by selecting the top row of the column and typing a title.

2. Press the **menu** key and select **4:Statistics** then select **1:Stat Calculations** then **1:One-Variable Statistics...**
3. Press **OK**
4. For **X1 List,** select the title for the column containing your data from the drop-down menu
5. Press **OK**

Note: You may need to scroll to view all of the statistics. This procedure also produces the mean, standard deviation, minimum, median, maximum.

IQR

JMP

1. Input the raw data into a column
2. Click **Analyze** and select **Distribution**
3. Click and drag the column name containing the data from the box under **Select Columns** to the box next to **Y, Columns**
4. Click **OK**
5. Click the red arrow next to **Summary Statistics** then select **Customize Summary Statistics**
6. Click the box next to Interquartile Range
7. Click **OK**

Minitab

1. Input the raw data into C1
2. Select **Stat** and choose **Basic Statistics** then choose **Display Descriptive Statistics...**
3. Double-click C1 to add it to the **Variables** list
4. Click the **Statistics** button
5. Check the box next to **Interquartile range**
6. Click **OK**
7. Click **OK**

SPSS

1. Input the raw data into the first column
2. Select **Analyze** and choose **Descriptive Statistics** then choose **Explore...**
3. Highlight the column name for the variable
4. Click the arrow to move the variable to the **Dependent List** box
5. Click **OK**

Note: These commands also produce the following statistics: mean, median, variance, standard deviation, minimum, maximum, and several plots.

Excel

1. Use the steps under the **Quartiles** section to find both the first and third quartiles
2. Click on an empty cell where you would like the result for iqr to appear
3. Type = into the cell
4. Click on the cell containing the third quartile
5. Type −
6. Click on the cell containing the first quartile
7. Press **Enter**

TI-83/84

The TI-83/84 does not have the functionality to produce the iqr automatically.

TI-Nspire

The TI-Nspire does not have the functionality to produce the iqr automatically.

Boxplot

JMP

1. Input the raw data into a column
2. Click **Analyze** and select **Distribution**
3. Click and drag the column name containing the data from the box under **Select Columns** to the box next to **Y, Columns**
4. Click **OK**

Minitab

1. Input the raw data into C1
2. Select **Graph** then choose **Boxplot...**
3. Highlight Simple under One Y
4. Click **OK**
5. Double-click C1 to add it to the **Graph Variables** box
6. Click **OK**

Note: You may add or format titles, axis titles, legends, and so on by clicking on the **Labels...** button prior to performing Step 6 above.

SPSS

1. Enter the raw data the first column
2. Select **Graph** and choose **Chart Builder...**
3. Under **Choose from** highlight Boxplot
4. Click and drag the first boxplot (Simple Boxplot) to the Chart preview area
5. Click and drag the data variable into the **Y-Axis?** box in the chart preview area
6. Click **OK**

Note: Boxplots are also produced when summary statistics such as mean, median, standard deviation, and so on are produced.

Excel

1. Input the raw data into the first column
2. Select the data (including title if used)
3. Click on the **Insert** Ribbon
4. Choose **Insert Statistic Chart** and select **Box and Whisker**
5. The chart will appear on the same worksheet as your data

Note: You may add or format titles, axis titles, legends, and so on, by right-clicking on the appropriate piece of the chart.

TI-83/84

1. Input the raw data into **L1** (To access lists press the **STAT** key, highlight the option called **Edit...** then press **ENTER**)
2. Press the **2nd** key and then press the **Y** = key
3. Select **Plot1** and press **ENTER**
4. Highlight **On** and press **ENTER**
5. Highlight the graph option in the second row, second column, and press **ENTER**
6. Press **ZOOM** then select **9:ZoomStat** to automatically resize the window and display the graph

TI-Nspire

1. Enter the data into a data list (To access data list select the spreadsheet option and press **enter**)

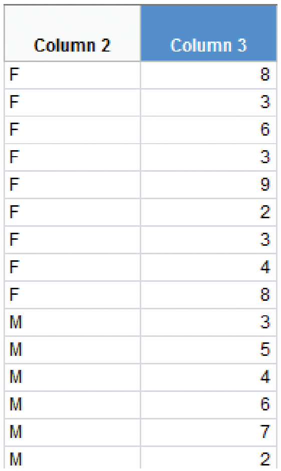
   Note: Be sure to title the list by selecting the top row of the column and typing a title.

2. Press **menu** and select **3:Data** then select **9:Quick Graph**
3. Press **menu** and select **1:Plot Type** then select **2:Box Plot** and press **enter**

Side-by-side Boxplots

JMP

1. Enter the raw data for both groups into a column
2. Enter the group information into a second column



Column 2	Column 3
F	8
F	3
F	6
F	3
F	9
F	2
F	3
F	4
F	8
M	3
M	5
M	4
M	6
M	7
M	2

Source: JMP

3. Click **Analyze** then select **Fit Y by X**
4. Click and drag the column name containing the raw data from the box under **Select Columns** to the box next to **Y, Response**

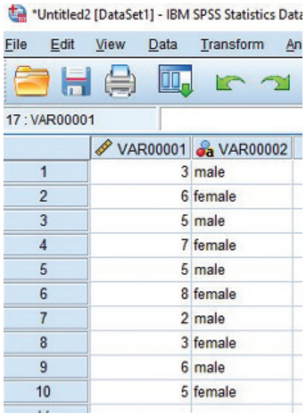
5. Click and drag the column name containing the group information from the box under **Select Columns** to the box next to **X, Factor**
6. Click **OK**
7. Click the red arrow next to **Oneway Analysis of...**
8. Click **Quantiles**

Minitab

1. Input the raw data for the first group into C1
2. Input the raw data for the second group into C2
3. Continue to input data for each group into a separate column
4. Select **Graph** then choose **Boxplot...**
5. Highlight Simple under Multiple Y's
6. Click **OK**
7. Double-click the column names for each column to be graphed to add it to the **Graph Variables** box
8. Click **OK**

SPSS

1. Enter the raw data into the first column
2. Enter the group data into the second column



17 : VAR00001

	VAR00001	VAR00002
1	3	male
2	6	female
3	5	male
4	7	female
5	5	male
6	8	female
7	2	male
8	3	female
9	6	male
10	5	female

Source: SPSS

3. Select **Graph** and choose **Chart Builder...**
4. Under **Choose from** highlight Boxplot
5. Click and drag the first boxplot (Simple Boxplot) to the Chart preview area
6. Click and drag the data variable into the **Y-Axis?** box in the chart preview area
7. Click and drag the group variable into the **X-Axis?** box in the chart preview area
8. Click **OK**

Excel

1. Enter the raw data into the first column
2. Enter the group data into the second column
3. Select an empty cell
4. Click on the **Insert** ribbon
5. Choose **Insert Statistic Chart** and select **Box and Whisker**
6. Right-click on the empty chart area that appears and select **Select Data...**
7. Under **Legend Entries (Series)** click **Add**
8. In the **Series name:** box select the column title for the data
9. In the **Series values:** box select the raw data values
10. Click **OK**
11. Under **Horizontal (Category) Axis Labels** click **Edit**
12. In the **Axis label range:** select the group labels (do NOT select the column title).
13. Click **OK**
14. Click **OK**

Note: You may add or format titles, axis titles, legends, and so on, by right-clicking on the appropriate piece of the chart.

TI-83/84

The TI-83/84 does not have the functionality to create side-by-side boxplots.

TI-Nspire

The TI-Nspire does not have the functionality to create side-by-side boxplots.

Copyright 2025 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s). Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

Chapter 5

Summarizing Bivariate Data

Luca Lanfranchi/Shutterstock.com

Elephant seals are very large marine mammals that spend most of their time in the ocean. However, each winter, they leave the water and spend about a month on the beach to mate and for the females to give birth to the pups that they have been carrying for a year. Body weight is one of the variables that biologists study in order to monitor the health of elephant seals.

It isn't easy to actually weigh an elephant seal (they are really big!). Because elephant seals are so difficult to weigh, researchers have considered ways to estimate weight so that they don't have to actually weigh them. In a study described in the paper **"Calibration of Aerial Photogrammetry to Estimate Elephant Seal Mass" (*Marine Mammal Science* [2020])**, researchers weighed 44 adult female elephant seals and also took pictures of the seals. They used the pictures to calculate a measure of size that they called the "footprint" of the seal. By studying the relationship between weight and footprint determined from a picture, they hoped to be able to estimate the weight of an elephant seal from its footprint.

In this chapter, you will see how a line that summarizes the relationship between two numerical variables can be used to make predictions, and what can be said about the accuracy of those predictions.

Learning Objectives

After completing this chapter, you will be able to:

- Informally describe the form, direction, and strength of a linear relationship based on a scatterplot.
- Calculate and interpret the value of the correlation coefficient.
- Explain the difference between a statistical relationship and a causal relationship (the difference between correlation and causation).
- Find the equation of the least-squares line and interpret the slope and intercept in context.
- Make predictions using the least-squares line.
- Calculate and interpret the value of the standard deviation about the least-squares line, s_e.
- Calculate and interpret the value of r^2, the coefficient of determination.
- Explain why it is important to consider both the standard deviation about the least-squares line and the value of r^2 when assessing the usefulness of the least-squares line.

Copyright 2025 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s). Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

- Construct a residual plot.
- Assess whether using a line to describe the relationship between two numerical variables is appropriate using a residual plot.
- Identify outliers and potentially influential observations in a linear regression context.

Section 5.1 Correlation

Sometimes statistical studies are carried out to investigate how two or more variables are related to one another. For example, an environmental researcher might want to know how the lead content of soil varies with distance from a major highway. Educational psychologists might wonder how vocabulary size is related to age. College admissions officers, who must try to predict whether an applicant will succeed in college, might be interested in the relationship between college grade point average and high school grade point average or score on an entrance exam.

As you saw in Chapter 3, you can get a visual impression of how strongly two numerical variables are related by looking at a scatterplot. It is also possible to describe the strength of a relationship numerically. The **correlation coefficient** (from *co-* and *relation*) is a numerical measure of the strength of the linear relationship between the x and y values in a bivariate data set consisting of (x, y) pairs.

Figure 5.1 displays several scatterplots that show different relationships between the x and y values.

What Do You Think?

Which of the scatterplots in Figure 5.1 shows the strongest relationship between x and y?

The plot in Figure 5.1(a) suggests a strong *positive relationship* between x and y. For every pair of points in the plot, the one with the larger x value also has the larger y value.

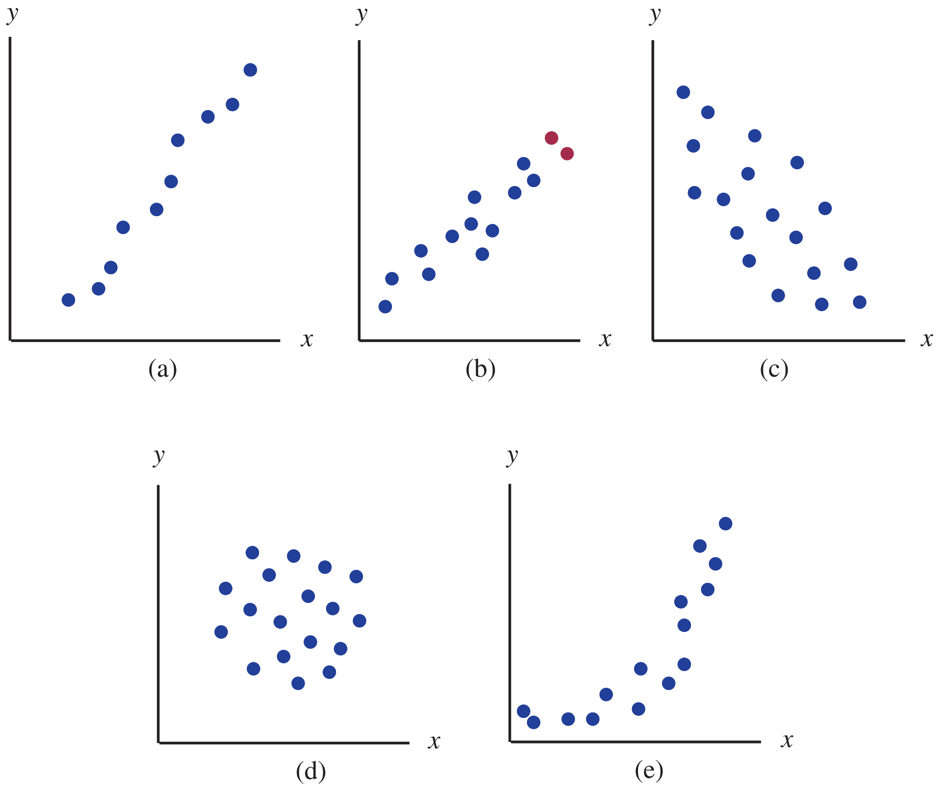


Figure 5.1 Scatterplots illustrating various types of relationships:
(a) positive linear relationship;
(b) positive linear relationship;
(c) negative linear relationship;
(d) no relationship;
(e) curved relationship.

Copyright 2025 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s). Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

The plot in Figure 5.1(b) shows a strong tendency for y to increase as x does, but there are a few exceptions. For example, the x and y values of the two points with the largest x values (shown in a different color) go in opposite directions (for this pair of points, x increases but y decreases in value). Even so, a plot like this still indicates a fairly strong positive relationship.

Figure 5.1(c) suggests that x and y are *negatively related*. As x increases, y tends to decrease. The negative relationship in this plot is not as strong as the positive relationship in Figure 5.1(b), although both plots show a well-defined linear pattern.

The plot of Figure 5.1(d) indicates that there is not a strong relationship between x and y. There is no tendency for y either to increase or to decrease as x increases. Finally, as illustrated in Figure 5.1(e), a scatterplot can show evidence of a strong relationship that is curved rather than linear.

The Sample Correlation Coefficient

The **sample correlation coefficient** measures the strength and direction of a linear relationship between two numerical variables. The correlation coefficient can be calculated using z scores. Consider replacing each x value by the corresponding z score, z_x (by subtracting $\bar{x}$ and then dividing by s_x) and similarly replacing each y value by its z score. Notice that x values that are larger than $\bar{x}$ will have positive z scores and those smaller than $\bar{x}$ will have negative z scores. Also y values larger than $\bar{y}$ will have positive z scores and those smaller will have negative z scores. The value of the correlation coefficient is based on the sum of the products of z_x and z_y for each observation in the bivariate data set. This can be written as $\Sigma z_x z_y$.

To see how this works, take a look at some scatterplots. The scatterplot in Figure 5.2(a) indicates a strong positive relationship. A vertical line through $\bar{x}$ and a horizontal line through $\bar{y}$ divide the plot into four regions. In Region I, both x and y are greater than their mean values, so the z score for x and the z score for y are both positive numbers. It follows that $z_x z_y$ is positive. The product of the z scores is also positive for any point in Region III, because both z scores are negative in Region III and multiplying two negative numbers results in a positive number. In each of the other two regions, one z score is positive and the other is negative, so $z_x z_y$ is negative. But because the points generally fall in Regions I and III, the products of z scores tend to be positive. This means that the *sum* of the products will be a relatively large positive number.

Similar reasoning for the data displayed in Figure 5.2(b), which exhibits a strong negative relationship, implies that $\Sigma z_x z_y$ will be a relatively large (in magnitude) negative number. When there is no strong relationship, as in Figure 5.2(c), positive and negative products tend to offset one another, producing a value of $\Sigma z_x z_y$ that is close to zero.

For these reasons, $\Sigma z_x z_y$ can be used as a measure of the strength of the linear relationship between x and y. It can be a large positive number, a large negative number, or a number close to 0, depending on whether there is a strong positive, a strong negative, or no linear relationship.

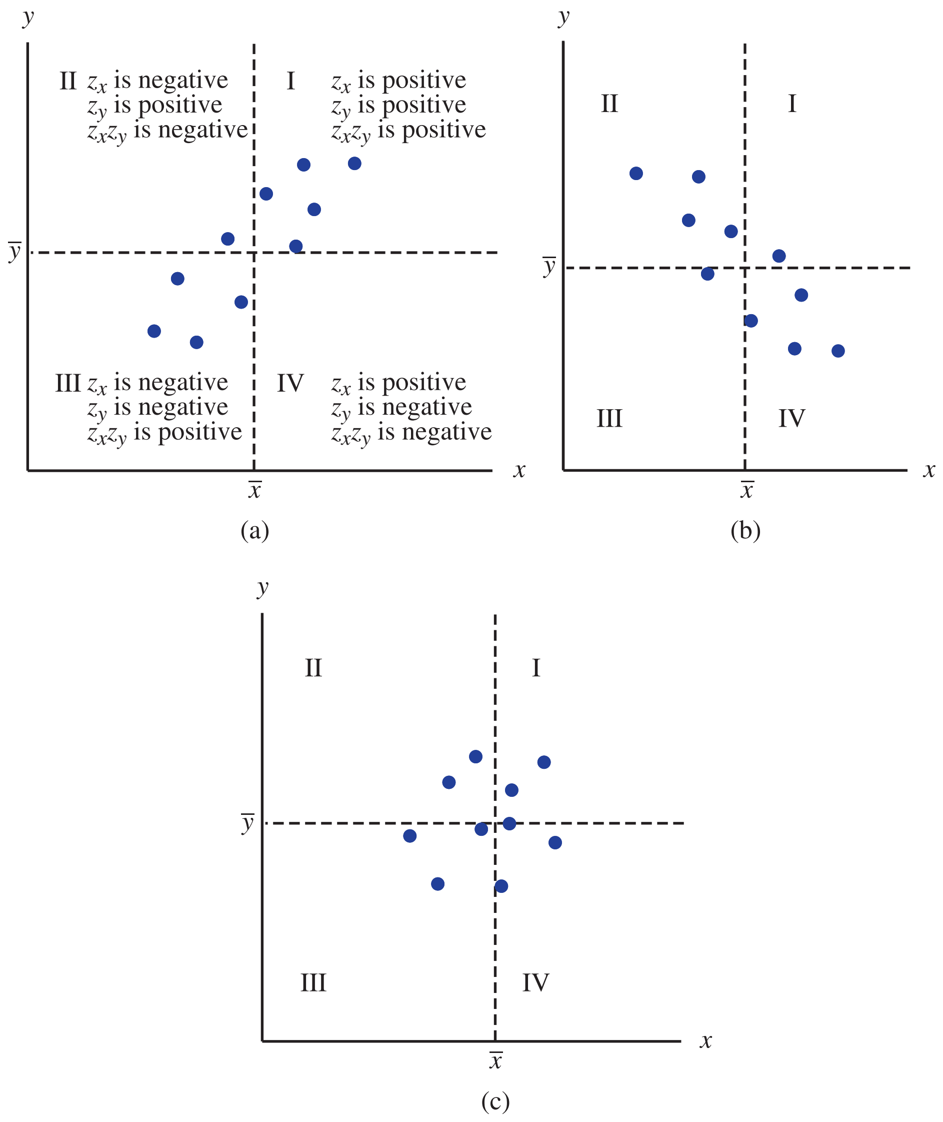
The value of the sample correlation coefficient, denoted by r, is calculated by dividing $\Sigma z_x z_y$ by $(n - 1)$.

Definition

Sample correlation coefficient: A measure of the strength and direction of a linear relationship between two numerical variables. Denoted by r, the sample correlation coefficient is given by

$$r = \frac{\Sigma z_x z_y}{n - 1}$$

Hand calculation of the correlation coefficient is quite tedious. Fortunately, all statistical software packages and most graphing calculators can compute the value of r using the x and y values.



(a)

(b)

(c)

Figure 5.2
Viewing a scatterplot according to the signs of z_x and z_y:
(a) a positive relationship;
(b) a negative relationship;
(c) no strong relationship.

Example 5.1 Long-Term Effects of Childhood Lead Exposure

Understand the context ➤

● The article **"Lead Exposure During Childhood Has Long-Lasting Effects" (The Economist, July 19, 2021)** summarizes data from a study of the relationship between childhood lead exposure and personality development. Participants in the study were adults in America and Europe. Researchers used self-assessment personality questionnaires to measure personality traits, including agreeableness and extraversion. These personality trait measures were then paired with information on atmospheric-lead pollution based on the postal code of a participant's childhood home. The accompanying table shows the average agreeableness score (on a scale from 1 to 5, with higher scores indicating more agreeable) and the average childhood atmospheric lead level for 10 countries in Europe (a subset of the data from a graph appearing in the article). The averages were calculated based on a sample of 10 participants from each country.

● **Data set available online**

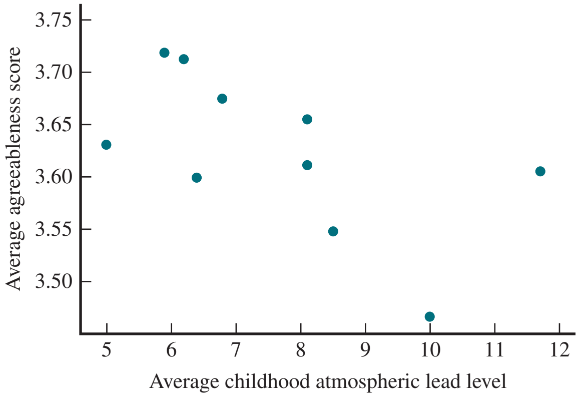
Copyright 2025 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s). Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

Consider the data ➤

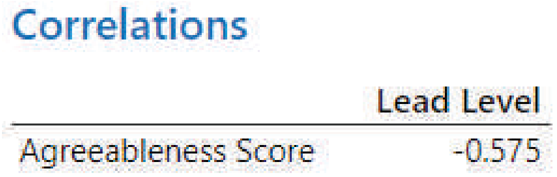
Atmospheric Lead Level (nanograms per cubic liter)	**Average Agreeableness Score**
5.0	3.63
6.8	3.67
8.1	3.66
10.0	3.47
11.7	3.61
5.9	3.72
6.4	3.60
8.5	3.55
6.2	3.71
8.1	3.61

Figure 5.3 is a scatterplot of these data. Based on the scatterplot, it appears that there is a negative relationship between agreeableness and childhood lead exposure that is approximately linear.

Figure 5.3
Scatterplot of agreeableness versus lead exposure.



Do the work ➤ With readily available technology, you will probably want to use a statistics software package or a graphing calculator to calculate the value of the correlation coefficient. Output from Minitab is shown here.



Correlations

	Lead Level
Agreeableness Score	-0.575

If you are calculating the value of the correlation coefficient by hand, it is a good idea to keep several decimal places in the intermediate calculations to avoid compounding rounding errors. In this example, three decimal places are used in the intermediate calculations.

Using x to denote childhood lead level and y to denote average agreeableness score, you can calculate the following values:

$$\bar{x} = 7.670 \qquad s_x = 2.042 \qquad \bar{y} = 3.623 \qquad s_y = 0.075$$

To calculate the value of the correlation coefficient, begin by calculating z-scores for each (x, y) pair in the data set. For example, the first observation is (5.0, 3.63). The corresponding z-scores are

$$z_x = \frac{5.0 - 7.670}{2.042} = -1.308 \qquad z_y = \frac{3.63 - 3.623}{0.075} = 0.093$$

The following table shows the z-scores and the product $z_x z_y$ for each observation.

x	y	z_x	z_y	$z_x z_y$
5.0	3.63	−1.308	0.093	−0.122
6.8	3.67	−0.426	0.627	−0.267
8.1	3.66	0.211	0.493	0.104
10.0	3.47	1.141	−2.040	−2.328
11.7	3.61	1.974	−0.173	−0.342
5.9	3.72	−0.867	1.293	−1.121
6.4	3.60	−0.622	−0.307	0.191
8.5	3.55	0.406	−0.973	−0.396
6.2	3.71	−0.720	1.160	−0.835
8.1	3.61	0.211	−0.173	−0.037

Then

$$\Sigma z_x z_y = -5.152$$

and

$$r = \frac{\Sigma z_x z_y}{n - 1} = \frac{-5.152}{10 - 1} = \frac{-5.152}{9} = -0.572$$

The small difference in the value calculated by Minitab (−0.575) and the value above (−0.572), is due to rounding in the hand calculations.

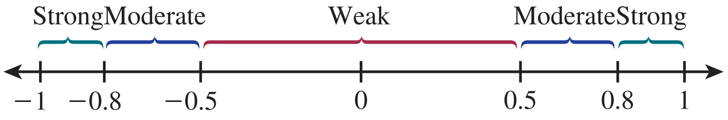
Interpret the results ➤

Based on the scatterplot and the value of the correlation coefficient, you would conclude that for these 10 countries, there is a moderate negative linear relationship between childhood atmospheric lead level and average agreeableness score. This is consistent with the conclusion given in the article referenced earlier, which says that based on the data for all 37 counties, adults in countries with higher levels of atmospheric lead had, on average, lower agreeableness scores. However, because this was an observational study, you should not conclude that higher lead exposure causes people to be less agreeable, since there may be potential confounding variables.

Properties of *r*

1. **The value of r is between -1 and $+1$.** A value near the upper limit, +1, indicates a strong positive linear relationship, whereas a value of r close to the lower limit, −1, suggests a strong negative linear relationship. Figure 5.4 shows a useful way to describe the strength of relationship based on r. It may seem surprising that a value of r as small as −0.5 or as large as 0.5 is in the weak category. An explanation for this is given later in the chapter.
2. **A correlation coefficient of $r = 1$ occurs only when all the points in a scatterplot of the data lie exactly on a straight line that slopes upward. Similarly, $r = -1$ only when all the points lie exactly on a downward-sloping line.** $r = 1$ or $r = -1$ indicates a perfect linear relationship between x and y in the sample data.

Figure 5.4
Describing the strength of a linear relationship.



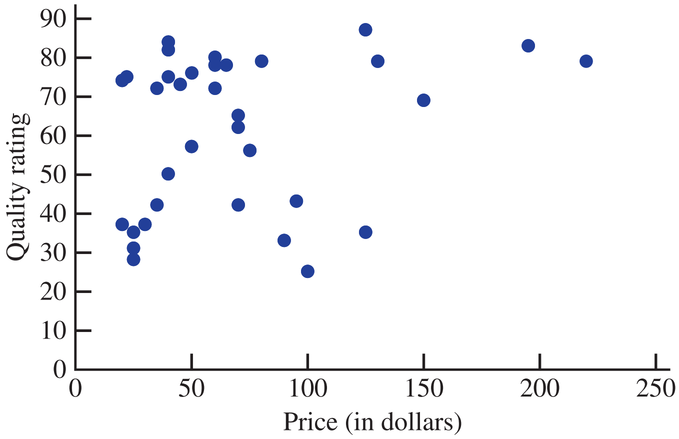
3. **The value of r is a measure of the extent to which x and y are linearly related.** r measures the extent to which the points in the scatterplot fall close to a straight line. A value of r close to 0 does not necessarily mean that there is *no* relationship between x and y. It is possible that there could still be a strong relationship that is not linear.
4. **The value of r does not depend on the unit of measurement for either variable.** For example, if x is height, the corresponding z score is the same whether height is expressed in inches, meters, or miles, and the value of the correlation coefficient is not affected. The correlation coefficient measures the inherent strength of the linear relationship between two numerical variables.
5. **The value of r does not depend on which of the two variables is considered to be the variable labeled as x.** This means that if you had let x = average agreeableness score and y = childhood lead level in Example 5.1, the same value, $r = -0.575$, would have resulted.

Example 5.2 Does It Pay to Pay More for a Bike Helmet?

Understand the context ➤

● Are more expensive bike helmets safer than less expensive ones? Data on x = price and y = quality rating for 35 different brands of bike helmets were used to construct the scatterplot in Figure 5.5. The data are from the ***Consumer Reports*** website **(consumerreports.org)**. Quality rating was a number from 0 (the worst possible rating) to 100, and was determined based on factors that included how well the helmet absorbed the force of an impact, the strength of the helmet, ventilation, and ease of use.

Figure 5.5
Minitab scatterplot for the bike helmet data of Example 5.2.



From the scatterplot, it appears that there is only a weak positive relationship between price and quality rating. The correlation coefficient, obtained using Minitab, is

Correlations: Price, Quality Rating

Correlation of Price and Quality Rating = 0.221

Interpret the results ➤

A correlation coefficient of $r = 0.221$ indicates that although there is a weak tendency for higher quality ratings to be associated with higher priced helmets, the relationship is not very strong. In fact, some inexpensive helmets had high quality ratings.

● **Data set available online**

Example 5.3 Age and Marathon Times

Understand the context ➤

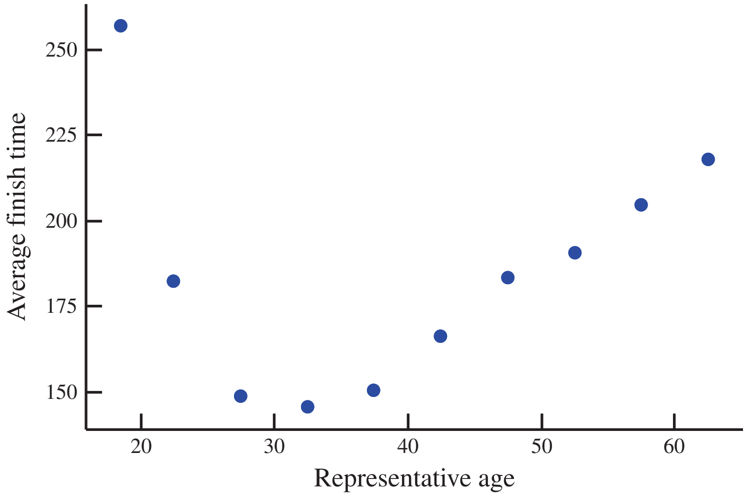
● Data from the New York Marathon website **(results.nyrr.org)** were used to calculate the average finishing time (in minutes) for the top 5 finishers by age group for participants who entered as women in the 2022 New York City marathon.

Consider the data ➤

Age Group	Representative Age	Average Finish Time of Top 5 (in minutes)
18–19	18.5	257.4
20–24	22.5	182.6
25–29	27.5	149.2
30–34	32.5	145.8
35–39	37.5	150.8
40–44	42.5	166.6
45–49	47.5	183.6
50–54	52.5	191.0
55–59	57.5	205.0
60–64	62.5	218.4

The scatterplot of average finish time versus representative age is shown in Figure 5.6.

Figure 5.6
Scatterplot of y = Average finish time and x = Representative age for the data of Example 5.3.



Using Minitab to compute the correlation coefficient between age and average finish time results in the following:

Correlations: Age, Average Finish Time

Correlation of Representative Age and Average Finish Time = 0.091

Interpret the results ➤

This example shows the importance of interpreting r as a measure of the strength of a *linear* relationship. Here, the value of r is small, but there is a strong nonlinear relationship between age and average finish time. This is an important point—you should not conclude that there is no relationship whatsoever simply because the value of r is small in absolute value. Be sure to look at the scatterplot of the data before concluding that there is no relationship between two variables based on a correlation coefficient with a value near 0.

● **Data set available online**

The Population Correlation Coefficient

The sample correlation coefficient r measures how strongly the x and y values in a *sample* of pairs are linearly related to one another. A measure of how strongly x and y are related in the entire population of pairs from which the sample was selected is called the **population correlation coefficient** and is denoted by ρ. (Notice again the use of a Greek letter for a population characteristic and a Roman letter for a sample characteristic.) It is unusual to calculate ρ from an entire population of pairs, but it is important to know that ρ satisfies properties paralleling those of r:

1. ρ is a number between -1 and $+1$ that does not depend on the unit of measurement for either x or y, or on which variable is labeled x and which is labeled y.
2. $\rho = +1$ or -1 if and only if all (x, y) pairs in the population lie exactly on a straight line, so ρ measures the extent to which there is a linear relationship in the population.

In Chapter 13, you will see how the sample correlation coefficient r can be used to draw conclusions about the value of the population correlation coefficient ρ.

Correlation and Causation

A value of r close to 1 indicates that the larger values of one variable tend to be associated with the larger values of the other variable. This is different from saying that a large value of one variable *causes* the value of the other variable to be large. Correlation measures the extent of association, but association *does not* imply causation.

It frequently happens that two variables are highly correlated not because one is causally related to the other but because they are both strongly related to a third variable. Among all elementary school children, the relationship between the number of cavities in a child's teeth and the size of their vocabulary is strong and positive. Yet no one advocates eating foods that result in more cavities to increase vocabulary size (or working to decrease vocabulary size to protect against cavities). Number of cavities and vocabulary size are both strongly related to age, so older children tend to have higher values of both variables than do younger ones. The article **"Study Suggests Bald Men Could be More At Risk For Severe Corona Virus, But More Research is Needed" (forbes.com, June 6, 2020)** describes several studies linking hair loss in men to severity of corona virus symptoms, but points out that the study did not control for age. Age is related to both hair loss and severity of symptoms, and this might account for the observed correlation between hair loss and severity.

Scientific experiments can frequently make a strong case for causality by carefully controlling the values of all variables that might be related to the ones under study. Then, if y is observed to change in a "smooth" way as the experimenter changes the value of x, a plausible explanation would be that there is a causal relationship between x and y. In the absence of such control and the ability to manipulate values of one variable, you must admit the possibility that an unidentified underlying third variable is influencing both variables under investigation. A high correlation in many uncontrolled studies carried out in different settings can marshal support for causality—as in the case of cigarette smoking and cancer—but establishing causality is an elusive task.

Exercises 5.1–5.16

● Data set available online

5.1 For each of the scatterplots shown on the next page, answer the following questions:

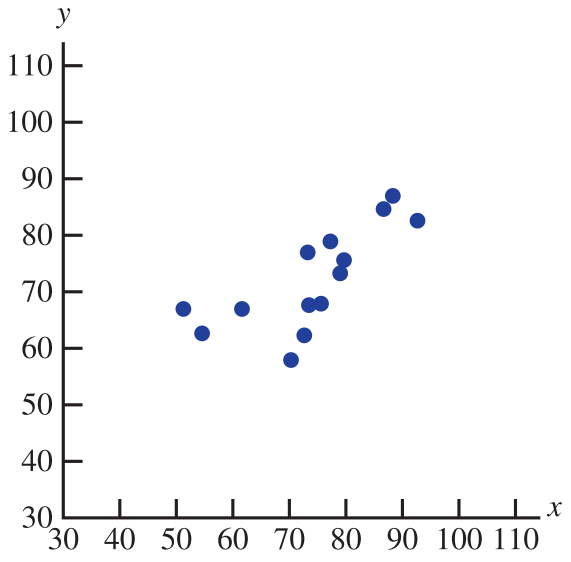
a. Does there appear to be a relationship between x and y?

b. If so, does the relationship appear to be linear?

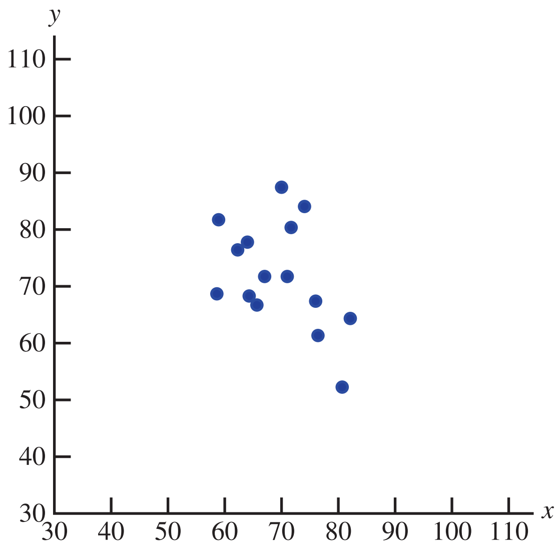
c. If so, would you describe the linear relationship as positive or negative?

Copyright 2025 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s). Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

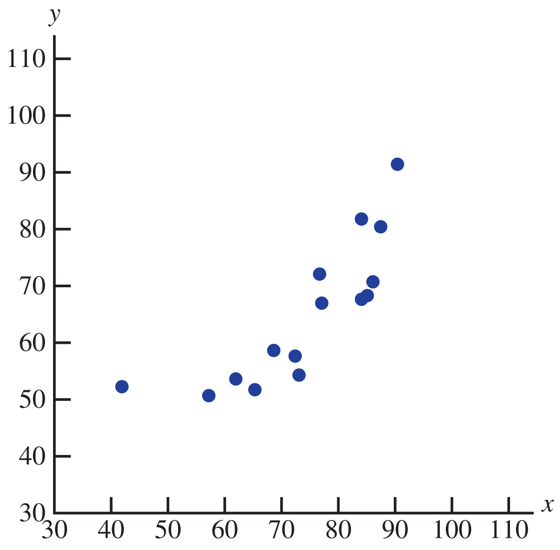
Scatterplot 1:



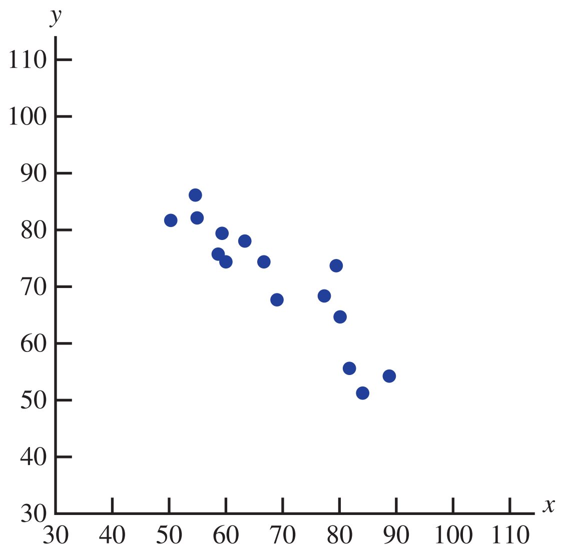
Scatterplot 2:



Scatterplot 3:



Scatterplot 4:



5.2 For each of the following pairs of variables, indicate whether you would expect a positive correlation, a negative correlation, or a correlation close to 0. Explain your choice.

a. Maximum daily temperature and cooling costs
b. Interest rate and number of loan applications
c. Amount of fertilizer used per acre and crop yield (Hint: As the amount of fertilizer is increased, yield tends to increase for a while but then tends to start decreasing.)

5.3 For each of the following pairs of variables, indicate whether you would expect a positive correlation, a negative correlation, or a correlation close to 0. Explain your choice.

a. Height and age in children age 2 to 12
b. Height and score on a college entrance exam
c. Height and shoe size

5.4 For each of the following pairs of variables, indicate whether you would expect a positive correlation, a negative correlation, or a correlation close to 0. Explain your choice.

a. Score on the math section of the SAT exam and score on the verbal section of the same test
b. Time spent on homework and time spent watching television during the same day by elementary school children

5.5 Is the following statement correct? Explain why or why not. (Hint: See Example 5.3.)

> A correlation coefficient of 0 implies that no relationship exists between the two variables under study.

5.6 Draw a scatterplot for which $r = 1$.

5.7 Draw a scatterplot for which $r = -1$.

Copyright 2025 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s). Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

5.8 ● Each year J.D. Power and Associates surveys new car owners 90 days after they purchase their cars. These data are used to rate auto brands (such as Toyota and Ford) on quality and customer satisfaction. Data for a quality rating and a satisfaction rating for 32 brands sold in the United States are given in the accompanying table.

Brand	Quality Rating	Satisfaction Rating
Acura	185	845
Audi	225	845
BMW	176	869
Buick	150	832
Cadillac	162	874
Chevrolet	141	839
Chrysler	189	828
Dodge	136	872
Ford	174	853
Genesis	142	864
GMC	151	857
Honda	177	842
Hyundai	153	842
Infiniti	173	846
Jaguar	190	854
Jeep	155	822
Kia	136	844
Land Rover	228	866
Lexus	159	859
Lincoln	182	876
Mazda	184	844
Mercedes-Benz	202	863
MINI	174	846
Mitsubishi	148	829
Nissan	161	844
Porsche	186	881
Ram	141	871
Subaru	187	832
Tesla	250	896
Toyota	177	825
Volkswagen	152	832
Volvo	210	859

a. Construct a scatterplot of y = Satisfaction rating versus x = Quality rating. How would you describe the relationship between x and y?

b. Calculate and interpret the value of the correlation coefficient.

5.9 ● The accompanying data are x = Cost (cents per serving) and y = Fiber content (grams per serving) for 18 high-fiber cereals rated by ***Consumer Reports*** **(consumerreports.org/health)**.

Cost per Serving	Fiber per Serving	Cost per Serving	Fiber per Serving
33	7	53	13
46	10	53	10
49	10	67	8
62	7	43	12
41	8	48	7
19	7	28	14
77	12	54	7
71	12	27	8
30	8	58	8

a. Calculate and interpret the value of the correlation coefficient for this data set. (Hint: See Example 5.1.)

b. The serving size differed for the different cereals, with serving sizes varying from $\frac{1}{2}$ cup to $1\frac{1}{4}$ cups. Converting price and fiber content to "per cup" rather than "per serving" results in the accompanying data. Is the correlation coefficient for the per cup data greater than or less than the correlation coefficient for the per serving data?

Cost per Cup	Fiber per Cup	Cost per Cup	Fiber per Cup
9.3	44.0	13.0	53.0
10.0	46.0	10.0	53.0
10.0	49.0	8.0	67.0
7.0	62.0	12.0	43.0
6.4	32.8	7.0	48.0
7.0	19.0	28.0	56.0
12.0	77.0	7.0	54.0
9.6	56.8	16.0	54.0
8.0	30.0	10.7	77.3

5.10 The authors of the paper **"Flat-footedness Is Not a Disadvantage for Athletic Performance in Children Aged 11 to 15 Years"** (***Pediatrics*** **[2009]: e386–e392)** studied the relationship between y = Arch height and scores on a number of different motor ability tests for 218 children. They reported the following correlation coefficients:

Motor Ability Test	Correlation between Test Score and Arch Height
Height of counter movement jump	−0.02
Hopping: average height	−0.10
Hopping: average power	−0.09
Balance, closed eyes, one leg	0.04
Toe flexion	0.05

Copyright 2025 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s). Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

a. Interpret the value of the correlation coefficient between average hopping height and arch height. What does the fact that the correlation coefficient is negative say about the relationship? Do higher arch heights tend to be paired with higher or lower average hopping heights?

b. The title of the paper suggests that having a small value for arch height (flat-footedness) is not a disadvantage when it comes to motor skills. Do the given correlation coefficients support this conclusion? Explain.

5.11 The paper **"The Relationship Between Cell Phone Use, Academic Performance, Anxiety, and Satisfaction with Life in College Students" (*Computers in Human Behavior* [2014]: 343–350)** described a study of cell phone use among undergraduate college students at a public university. The paper reported that the value of the correlation coefficient between x = Cell phone use (measured as total amount of time in hours spent using a cell phone on a typical day) and y = GPA (cumulative GPA determined from university records) was $r = -0.203$.

a. Interpret the given value of the correlation coefficient. Does the value of the correlation coefficient suggest that students who use a cell phone for more hours per day tend to have higher GPAs or lower GPAs?

b. The study also investigated the correlation between texting (measured as the total number of texts sent and texts received per day) and GPA. The direction of the relationship between texting and GPA was the same as the direction of the relationship between cell phone use and GPA, but the relationship between texting and GPA was not as strong. Which of the following possible values for the correlation coefficient between texting and GPA could have been the one observed?

$r = -0.30$ $r = -0.10$ $r = 0.10$ $r = 0.30$

c. The paper included the following statement: "Participants filled in two blanks—one for texts sent and one for texts received. These two texting items were nearly perfectly correlated." Do you think that the value of the correlation coefficient for texts sent and texts received was close to -1, close to 0, or close to $+1$? Explain your reasoning.

5.12 Several studies that are summarized in the paper **"Does Social Media Use Increase Depressive Symptoms? A Reverse Causation Perspective" (*Frontiers in Psychiatry*, 2021)** point to a link between social media use and depression.

a. If an observational study reveals that a higher rate of depression tends to occur in groups with greater social media use, would the value of the correlation coefficient be positive or negative?

b. The authors of the paper point out that based just on a correlation between rate of depression and social media use, it is not possible to determine if increased social media use causes an increased rate of depression, increased depression causes greater social media use, or there is no causal relationship between social media use and rate of depression. Do you agree? Explain why or why not.

5.13 ● The ability to jump helps salamanders move quickly to avoid predators. The paper **"Jumping in Arboreal Salamanders: A Possible Tradeoff Between Takeoff Velocity and In-Air Posture" (*Zoology* [2020]: 125724)** investigated jumping behavior of 10 species of salamanders and reported the following data on bend angle (the bend [in degrees] in the salamander body, with a larger number reflecting more extreme bending) and the jump takeoff velocity (in meters per second).

Bend Angle	Takeoff Velocity
36.5	0.6
55.0	0.7
43.5	0.7
50.4	0.7
69.8	0.9
71.2	1.0
67.4	1.1
48.1	0.7
61.8	0.9
52.6	0.9

a. Construct a scatterplot of y = Takeoff velocity versus x = Bend angle. How would you describe the relationship between bend angle and takeoff velocity?

b. Calculate and interpret the value of the correlation coefficient.

c. Explain how the value of the correlation coefficient is consistent with your description of the relationship in Part (a).

5.14 It may seem odd, but one of the ways biologists can tell how old a lobster is involves measuring the concentration of a pigment called neurolipofuscin in the eyestalk of a lobster. (We are not making this up!) The authors of the paper **"Neurolipofuscin Is a Measure of Age in *Panulirus argus*, the Caribbean Spiny Lobster, in Florida"**

Copyright 2025 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s). Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

(*Biological Bulletin* [2007]: 55–66) wondered if it was sufficient to measure the pigment in just one eye stalk, which would be the case if there is a strong relationship between the concentration in the right and left eyestalks.

Pigment concentration (as a percentage of tissue sample) was measured in both eyestalks for 39 lobsters, resulting in the following summary quantities (based on data read from a graph that appeared in the paper):

$n = 39$ $\Sigma x = 88.8$ $\Sigma y = 86.1$

$\Sigma xy = 281.1$ $\Sigma x^2 = 288.0$ $\Sigma y^2 = 286.6$

An alternative formula for calculating the correlation coefficient that is based on raw data and is algebraically equivalent to the one given in the text is

$$r = \frac{\Sigma xy - \dfrac{(\Sigma x)(\Sigma y)}{n}}{\sqrt{\Sigma x^2 - \dfrac{(\Sigma x)^2}{n}}\sqrt{\Sigma y^2 - \dfrac{(\Sigma y)^2}{n}}}$$

Use this formula to calculate the value of the correlation coefficient, and interpret this value.

5.15 An auction house released a list of 25 recently sold paintings. Eight artists were represented in these sales. The sale price of each painting also appears on the list. Would the correlation coefficient be an appropriate way to summarize the relationship between artist (x) and sale price (y)? Why or why not?

5.16 A sample of automobiles traversing a certain stretch of highway is selected. Each one travels at roughly a constant rate of speed, although speed does vary from auto to auto. Let x = speed and y = time needed to traverse this segment of highway. Would the sample correlation coefficient be closest to 0.9, 0.3, −0.3, or −0.9? Explain.

Section 5.2 Linear Regression: Fitting a Line to Bivariate Data

The objective of *regression analysis* is to use information about one variable, x, to predict the value of a second variable, y. For example, you might want to predict y = Product sales during a given period when the amount spent on advertising is x = \$10,000. The two variables in a regression analysis play different roles: y is called the **dependent** or **response variable**, and x is referred to as the **independent**, **predictor**, or **explanatory variable**.

Definitions

Dependent variable: In a bivariate data set, the variable whose value you would like to predict. The dependent variable is denoted by y. The dependent variable is also sometimes called the **response variable**.

Independent variable: In a bivariate data set, the variable that will be used to make a prediction of the dependent variable. The independent variable is denoted by x. The independent variable is also sometimes called the **predictor variable** or the **explanatory variable**.

Scatterplots frequently exhibit a linear pattern. When this is the case, it makes sense to summarize the relationship between the variables using a line. Before seeing how this is done, let's review some elementary facts about lines and linear relationships.

The equation of a line is $y = a + bx$. A particular line is specified by choosing values of a and b. For example, one line is $y = 10 + 2x$; another is $y = 100 - 5x$. If you choose some x values and calculate $y = a + bx$ for each value, the points in the plot of the resulting (x, y) pairs will fall exactly on a straight line.

Definitions

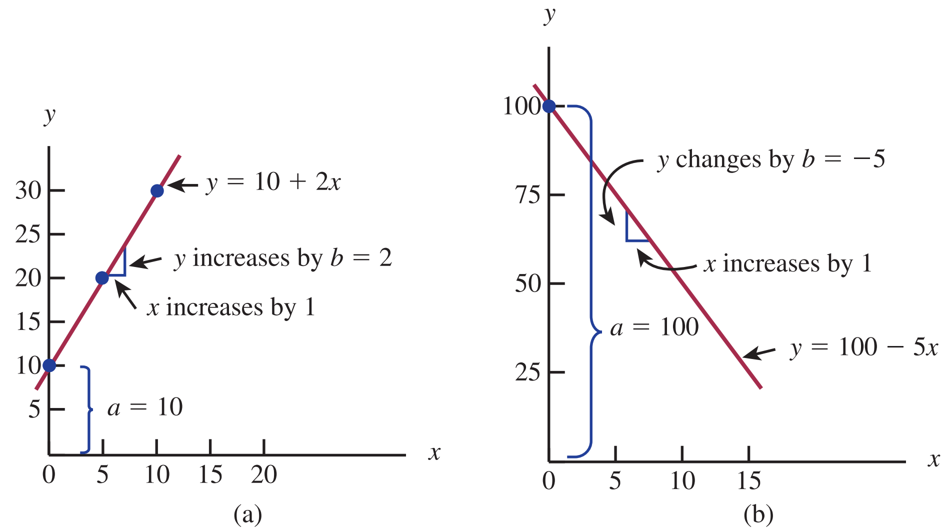
The equation of a line is

$$y = \overset{\text{Intercept}}{a} + \underset{\text{Slope}}{b}x$$

Slope: The value of b, called the **slope** of the line, is the amount by which y increases when x increases by 1 unit.

Intercept: The value of a, called the **intercept** (or sometimes the ***y*-intercept** or **vertical intercept**) of the line, is the height of the line above the value $x = 0$.

The line $y = 10 + 2x$ has slope $b = 2$, so each 1-unit increase in x is paired with an increase of 2 in y. When $x = 0$, $y = 10$, so the height of the line where it crosses the vertical axis (where $x = 0$) is 10. This is illustrated in Figure 5.7(a). The slope of the line $y = 100 - 5x$ is -5, so y increases by -5 (or equivalently, decreases by 5) when x increases by 1. The height of the line above $x = 0$ is $a = 100$. This line is pictured in Figure 5.7(b).



(a) (b)

Figure 5.7
Graphs of two lines:
(a) slope $b = 2$, intercept $a = 10$;
(b) slope $b = -5$, intercept $a = 100$.

It is easy to draw the graph of a line. Choose any two x values and substitute them into the equation of the line to obtain the two corresponding y values. Then plot the resulting two (x, y) pairs as two points. The line passes through these points. For the equation $y = 10 + 2x$, substituting $x = 5$ yields $y = 20$, and using $x = 10$ gives $y = 30$. The resulting two points are then (5, 20) and (10, 30). The line in Figure 5.7(a) passes through these points.

Fitting a Line: The Principle of Least Squares

Figure 5.8 shows a scatterplot that also includes two lines. Line II is a better fit to the data than Line I is. Vertical deviations from the line can be used to measure the extent to which a particular line provides a good fit to data.

Line II in Figure 5.8 has equation $y = 10 + 2x$. The third and fourth points from the left in the scatterplot are (15, 44) and (20, 45). For these two points, the vertical deviations from this line are

$$\begin{aligned}\text{3rd deviation} &= y_3 - \text{height of the line above } x_3\\ &= 44 - [10 + 2(15)]\\ &= 4\end{aligned}$$

and

$$\text{4th deviation} = 45 - [10 + 2(20)] = -5$$

Copyright 2025 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s). Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

What Do You Think?

The deviation for the fourth observation just calculated was negative. When will a deviation from a line be negative?

A positive vertical deviation results from a point that lies above the line, and a negative deviation results from a point that lies below the line.

A particular line is said to be a good fit to the data if the deviations from the line are small in magnitude. Line I in Figure 5.8 fits poorly because all deviations from that line are larger in magnitude (and some are much larger) than the corresponding deviations from Line II.

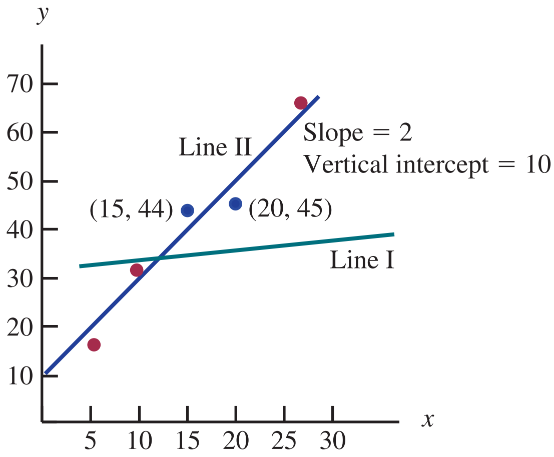


Figure 5.8
Line I gives a poor fit and Line II gives a good fit to the data.

To assess the overall fit of a line, a way to combine all of the deviations into a single measure of fit is needed. One way is to square the deviations (to obtain nonnegative numbers) and to then add up the squared deviations.

Definitions

Sum of squared deviations: The most widely used measure of the goodness of fit of a line $y = a + bx$ to bivariate data $(x_1, y_1), \dots, (x_n, y_n)$ is the **sum of the squared deviations** about the line

$$\Sigma[y - (a + bx)]^2 = [y_1 - (a + bx_1)]^2 + [y_2 - (a + bx_2)]^2 + \cdots + [y_n - (a + bx_n)]^2$$

Least-squares line: The line that minimizes the sum of squared deviations. The least-squares line is also sometimes called the **sample regression line**.

Fortunately, the equation of the least-squares line can be obtained without having to calculate deviations from any particular line. The accompanying box gives formulas for the slope and intercept of the least-squares line.

The slope of the least-squares line is

$$b = \frac{\Sigma(x - \bar{x})(y - \bar{y})}{\Sigma(x - \bar{x})^2}$$

and the y intercept is

$$a = \bar{y} - b\bar{x}$$

The equation of the least-squares line is written as

$$\hat{y} = a + bx$$

where the ˆ above y indicates that $\hat{y}$ (read as y-hat) is the prediction of y that results from substituting a particular x value into the equation.

Statistical software packages and many calculators can compute the slope and intercept of the least-squares line. If the slope and intercept are to be calculated by hand, the following calculating formula can be used to reduce the amount of time required to perform the calculations.

Calculating Formula for the Slope of the Least-Squares Line

$$b = \frac{\Sigma xy - \dfrac{(\Sigma x)(\Sigma y)}{n}}{\Sigma x^2 - \dfrac{(\Sigma x)^2}{n}}$$

Example 5.4 Pomegranate Juice and Tumor Growth

Understand the context ➤

● Pomegranate, a fruit native to Persia, has been used in the folk medicines of many cultures to treat various ailments. Researchers have studied pomegranate's antioxidant properties to see if it might have any beneficial effects in the treatment of cancer. One study, described in the paper **"Pomegranate Fruit Juice for Chemoprevention and Chemotherapy of Prostate Cancer" (*Proceedings of the National Academy of Sciences* [October 11, 2005]: 14813–14818)**, investigated whether pomegranate fruit extract (PFE) was effective in slowing the growth of prostate cancer tumors.

In this study, 24 mice were injected with cancer cells. The mice were then randomly assigned to one of three treatment groups. One group of eight mice received normal drinking water. The second group of eight mice received drinking water supplemented with 0.1% PFE and the third group received drinking water supplemented with 0.2% PFE. The average tumor volume for the mice in each group was recorded at several points in time.

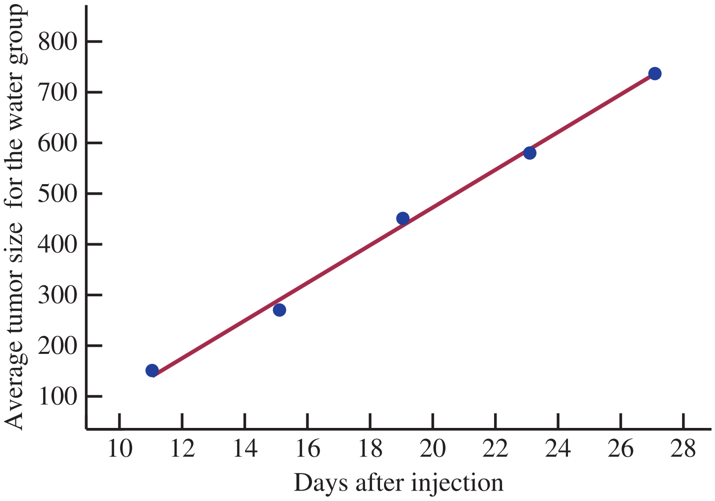
The accompanying data on y = Average tumor volume (in mm^3) and x = Number of days after injection of cancer cells for the mice that received plain drinking water were approximated from a graph that appeared in the paper:

Consider the data ➤

x	11	15	19	23	27
y	150	270	450	580	740

A scatterplot of these data (Figure 5.9) shows that the relationship between number of days after injection of cancer cells and average tumor volume could reasonably be described by a line.

Figure 5.9
Scatterplot for the plain drinking water data of Example 5.4.



● **Data set available online**

Copyright 2025 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s). Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

Calculating the values of the slope and intercept of the least-squares line can be very tedious. This is where a statistics computer package or a graphing calculator can really be helpful. Minitab was used to find the equation of the least-squares line, and the resulting output is shown here:

Regression Analysis: *y* versus *x*

```
The regression equation is
y = −270 + 37.3x

Predictor        Coef    SE Coef        T        P
Constant      −269.75      23.42   −11.52    0.001
x              37.250      1.181    31.53    0.000
```

From the Minitab output, the equation of the least-squares line is

$$\hat{y} = -270 + 37.3x$$

You can also use the values given in the Coef column for the slope and intercept. These values have more decimal accuracy than those in the given equation.

The values of the slope and intercept of the least-squares line can also be calculated by hand. The summary quantities necessary to calculate these values are

$$\Sigma x = 95 \qquad \Sigma x^2 = 1965 \qquad \Sigma xy = 47{,}570$$
$$\Sigma y = 2190 \qquad \Sigma y^2 = 1{,}181{,}900$$

From these quantities,

Do the work ➤ $\bar{x} = 19 \qquad \bar{y} = 438$

$$b = \frac{\Sigma xy - \dfrac{(\Sigma x)(\Sigma y)}{n}}{\Sigma x^2 - \dfrac{(\Sigma x)^2}{n}} = \frac{47{,}570 - \dfrac{(95)(2190)}{5}}{1965 - \dfrac{(95)^2}{5}} = \frac{5960}{160} = 37.25$$

and

$$a = \bar{y} - b\bar{x} = 438 - (37.25)(19) = -269.75$$

The equation of the least-squares line is then

$$\hat{y} = -269.75 + 37.25x$$

This line is also shown on the scatterplot of Figure 5.9.

If you wanted to predict average tumor volume 20 days after injection of cancer cells, you can use the y value of the point on the least-squares line above $x = 20$:

$$\hat{y} = -269.75 + 37.25(20) = 475.25$$

Predicted average tumor volume for other numbers of days after injection of cancer cells could be calculated in a similar way.

Sometimes you need to be cautious when making predictions. For example, in the context of Example 5.4, the least-squares line should not be used to predict average tumor volume for times much outside the range 11 to 27 days (the range of x values in the data set) because you do not know whether the linear pattern observed in the scatterplot continues outside this range. This is sometimes referred to as the **danger of extrapolation.**

Using the least-squares line to predict average tumor volume for fewer than 10 days after injection of cancer cells can lead to nonsensical predictions. For example, if the number of days after injection is five, the predicted average tumor volume is negative:

$$\hat{y} = -269.75 + 37.25(5) = -83.5$$

Because it is impossible for average tumor volume to be negative, this is a clear indication that the pattern observed for x values in the 11 to 27 range does not continue outside that range. However, the least-squares line can still be a useful tool for making predictions for x values within the 11- to 27-day range.

USE CAUTION—The Danger of Extrapolation

The least-squares line should not be used to make predictions outside the range of the x values in the data set because there is no evidence that the linear relationship continues outside this range.

Example 5.5 More on Pomegranate Juice and Tumor Growth

Figure 5.10 shows a scatterplot for average tumor volume versus number of days after injection of cancer cells for both the group of mice that drank only water and the group that drank water supplemented by 0.2% PFE. Notice that the tumor growth seems to be much slower for the mice that drank water supplemented with PFE. For the 0.2% PFE group, the relationship between average tumor volume and number of days after injection of cancer cells appears to be curved rather than linear. You will see in Section 5.4 how a curve (rather than a straight line) can be used to describe nonlinear relationships.

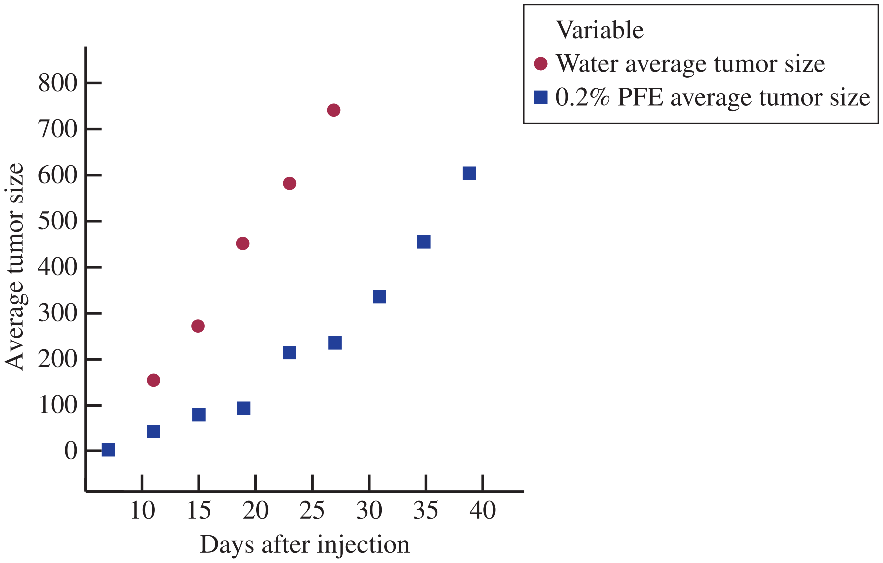


Figure 5.10 Scatterplot of average tumor volume versus number of days after injection of cancer cells for the water group and the 0.2% PFE group.

Example 5.6 Ankle Motion and Balance

Understand the context ➤

● Good balance is important in preventing falls, especially in older adults. The paper **"Correlation Between Ankle Range of Motion and Balance in Community-Dwelling Elderly Population" (*Indian Journal of Physiotherapy and Occupational Therapy* [2012]: 127–129)** describes a study of the relationship between ankle range of motion and balance. The accompanying data on x = Ankle range of motion (in degrees) and y = Balance test score for 30 people ages 65 to 85 are consistent with a scatterplot and summary measures given in the paper.

● **Data set available online**

Copyright 2025 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s). Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

Consider the data ➤

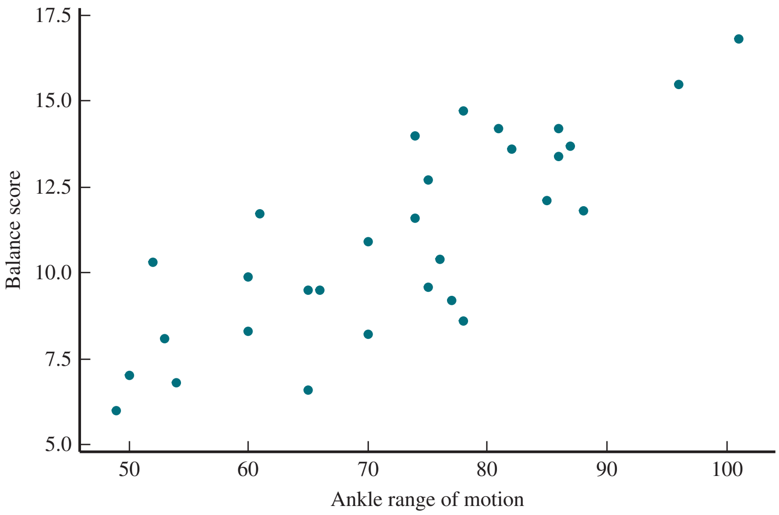
x = Ankle Range of Motion	y = Balance Score	x = Ankle Range of Motion	y = Balance Score
85	12.1	82	13.6
74	14.0	78	14.7
76	10.4	86	14.2
88	11.8	86	13.4
49	6.0	101	16.8
75	9.6	75	12.7
70	8.2	52	10.3
60	9.9	70	10.9
81	14.2	77	9.2
54	6.8	53	8.1
50	7.0	87	13.7
96	15.5	65	6.6
60	8.3	78	8.6
65	9.5	66	9.5
61	11.7	74	11.6

Minitab was used to construct a scatterplot of the data (Figure 5.11) and to compute the value of the correlation coefficient.

Correlations: Ankle Range of Motion, Balance Score

Pearson correlation of Ankle Range of Motion and Balance Score = 0.807 P-Value = 0.000

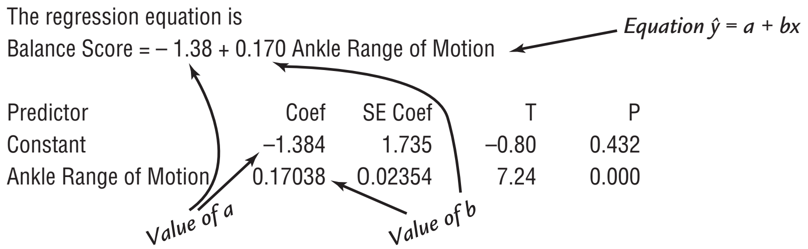
Figure 5.11
Minitab scatterplot for the data of Example 5.6.



Notice that there is a linear pattern in the scatterplot, suggesting that a line is a reasonable way to summarize the relationship between balance score and ankle range of motion. The correlation coefficient of $r = 0.807$ is consistent with the pattern in the scatterplot and indicates a strong positive linear relationship.

Figure 5.12 shows part of the Minitab regression output. Instead of x and y, the labels "Balance Score" and "Ankle Range of Motion" are used. In the table just below the equation of the regression line, the first row gives information about the intercept, a, and the second row gives information about the slope, b. Notice that the coefficient column, labeled "Coef," contains more precise values for a and b than those that appear in the equation.

Copyright 2025 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s). Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.



```
The regression equation is
Balance Score = – 1.38 + 0.170 Ankle Range of Motion

Predictor                  Coef    SE Coef      T       P
Constant                 –1.384      1.735  –0.80   0.432
Ankle Range of Motion   0.17038    0.02354   7.24   0.000
```

Equation $\hat{y} = a + bx$

Value of a

Value of b

Figure 5.12
Partial Minitab output for Example 5.6.

Figure 5.13 shows the scatterplot with the least squares regression line included. The equation of the least squares regression line

$$\hat{y} = -1.38 + 0.170x$$

can be used to predict the balance score for a given ankle range of motion. For example, for a person with ankle range of motion of 85 degrees, you would predict a balance score of

$$\hat{y} = -1.38 + 0.170(85) = 13.07$$

The least squares regression line should not be used to predict the balance score for people with ankle ranges of motion such as $x = 30$ or $x = 120$. These x values are well outside the range of the data, and you do not know if a linear relationship is appropriate outside the observed range.

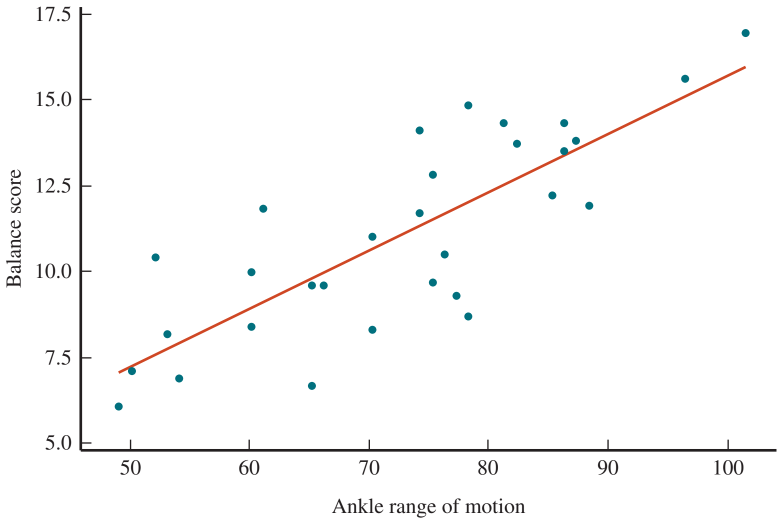


Figure 5.13
Scatterplot and least squares regression line for the range of motion and balance data of Example 5.6.

Regression

The least-squares line is often called the **sample regression line**. The term regression comes from the relationship between the least-squares line and the correlation coefficient. To understand this relationship, consider alternative expressions for the slope b and the equation of the line itself. With s_x and s_y denoting the sample standard deviations of the x's and y's, respectively, a bit of algebraic manipulation gives

$$b = r\left(\frac{s_y}{s_x}\right)$$

$$\hat{y} = \bar{y} + r\left(\frac{s_y}{s_x}\right)(x - \bar{x})$$

You do not need to use these formulas in any calculations, but several of their implications are important for appreciating what the least-squares line does.

1. When $x = \bar{x}$ is substituted in the equation of the line, $\hat{y} = \bar{y}$ results. This means that the least-squares line passes through the *point of averages* $(\bar{x}, \bar{y})$.
2. Suppose for the moment that $r = 1$, so that all points lie exactly on the line whose equation is

$$\hat{y} = \bar{y} + \frac{s_y}{s_x}(x - \bar{x})$$

Copyright 2025 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s). Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

Now substitute $x = \bar{x} + s_x$, which is 1 standard deviation above $\bar{x}$:

$$\hat{y} = \bar{y} + \frac{s_y}{s_x}(\bar{x} + s_x - \bar{x}) = \bar{y} + s_y$$

Notice that with $r = 1$, when x is 1 standard deviation above its mean, you would predict that the associated y value will be 1 standard deviation above its mean. Similarly, if $x = \bar{x} - 2s_x$ (2 standard deviations below its mean), then

$$\hat{y} = \bar{y} + \frac{s_y}{s_x}(\bar{x} - 2s_x - \bar{x}) = \bar{y} - 2s_y$$

which is also 2 standard deviations below the mean.

If $r = -1$, then $x = \bar{x} + s_x$ results in $\hat{y} = \bar{y} - s_y$, so the predicted y is also 1 standard deviation from its mean but on the opposite side of $\bar{y}$ from where x is relative to $\bar{x}$. In general, if x and y are perfectly correlated, the predicted y value associated with a given x value will be the same number of standard deviations (of y) from its mean $\bar{y}$ as x is from its mean $\bar{x}$.

3. Now suppose that x and y are not perfectly correlated. For example, suppose $r = 0.5$, so the least-squares line has the equation

$$\hat{y} = \bar{y} + 0.5\left(\frac{s_y}{s_x}\right)(x - \bar{x})$$

Then substituting $x = \bar{x} + s_x$ gives

$$\hat{y} = \bar{y} + 0.5\left(\frac{s_y}{s_x}\right)(\bar{x} + s_x - \bar{x}) = \bar{y} + 0.5s_y$$

For $r = 0.5$, when x lies 1 standard deviation above its mean, you would predict that y will be only 0.5 standard deviation above its mean. Similarly, you can predict y when r is negative. If $r = -0.5$, then the predicted y value will be only half the number of standard deviations from $\bar{y}$ that x is from $\bar{x}$ but x and the predicted y will now be on opposite sides of their respective means.

Consider using the least-squares line to predict the value of y associated with an x value some specified number of standard deviations away from $\bar{x}$. Then the predicted y value will be only r times this number of standard deviations from $\bar{y}$. In terms of standard deviations, except when $r = 1$ or -1, the predicted y will always be closer to $\bar{y}$ than x is to $\bar{x}$.

Using the least-squares line for prediction results in a predicted y that is pulled back in, or regressed, toward the mean of y compared to where x is relative to the mean of x. This regression effect was first noticed by Sir Francis Galton (1822–1911), a famous biologist, when he was studying the relationship between the heights of fathers and their sons. He found that predicted heights of sons whose fathers were above average in height were also above average (because r is positive here) but not by as much as the father's height. He found a similar relationship for fathers whose heights were below average. This regression effect has led to the term **regression analysis** for the collection of methods involving the fitting of lines, curves, and more complicated functions to bivariate and multivariate data.

The alternative form of the regression (least-squares) line emphasizes that predicting y from knowledge of x is not the same problem as predicting x from knowledge of y. The slope of the least-squares line for predicting x is $r(s_x/s_y)$ rather than $r(s_y/s_x)$ and the intercepts of the lines are also almost always different. For purposes of prediction, it makes a difference whether y is regressed on x, or x is regressed on y.

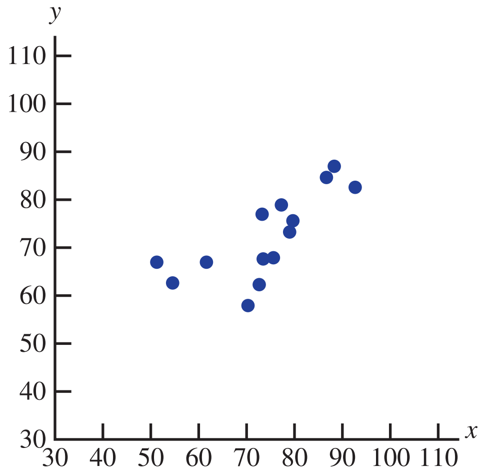
The least-squares line of y on x should not be used to predict x, because it is not the line that minimizes the sum of squared deviations in the x direction.

Exercises 5.17–5.34

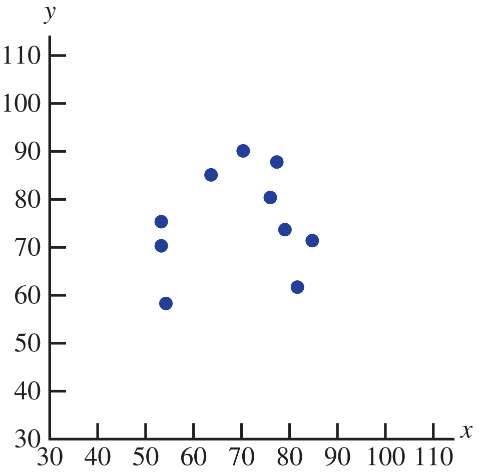
● **Data set available online**

5.17 Two scatterplots are shown below. Explain why it makes sense to use the least-squares line to summarize the relationship between x and y for one of these data sets but not the other. (Hint: See Example 5.5.)

Scatterplot 1:



Scatterplot 2:



5.18 ● The authors of the paper **"Statistical Methods for Assessing Agreement Between Two Methods of Clinical Measurement" (*International Journal of Nursing Studies* [2010]: 931–936)** compared two different instruments for measuring a person's ability to breathe out air. (This measurement is helpful in diagnosing various lung disorders.) The two instruments considered were a Wright peak flow meter and a mini-Wright peak flow meter. Seventeen subjects participated in the study, and for each person air flow was measured once using the Wright meter and once using the mini-Wright meter.

Subject	Mini-Wright Meter	Wright Meter	Subject	Mini-Wright Meter	Wright Meter
1	512	494	10	445	433
2	430	395	11	432	417
3	520	516	12	626	656
4	428	434	13	260	267
5	500	476	14	477	478
6	600	557	15	259	178
7	364	413	16	350	423
8	380	442	17	451	427
9	658	650			

a. Suppose that the Wright meter is considered to provide a better measure of air flow, but the mini-Wright meter is easier to transport and to use. If the two types of meters produce different readings but there is a strong relationship between the readings, it would be possible to use a reading from the mini-Wright meter to predict the reading that the larger Wright meter would have given. Use the given data to find an equation to predict Wright meter reading using a reading from the mini-Wright meter. (Hint: See Example 5.4.)

b. What would you predict for the Wright meter reading for a person whose mini-Wright meter reading was 500?

c. What would you predict for the Wright meter reading for a person whose mini-Wright meter reading was 300? (Hint: See the discussion of extrapolation that follows Example 5.4.)

5.19 ● The accompanying data are a subset of data from the report **"Great Jobs, Great Lives" (Gallup-Purdue Index 2015 Report, gallup.com)**. The values are approximate values read from a scatterplot. Students at a number of universities were asked if they agreed that their education was worth the cost. One variable in the table is the percentage of students at the university who responded *strongly agree*. The other variable in the table is the *U.S. News and World Report* ranking of the university.

Ranking	Percentage of Alumni Who Strongly Agree
28	53
29	58
30	62
37	55
45	54
47	62
52	55

(continued)

Ranking	Percentage of Alumni Who Strongly Agree
54	62
57	70
60	58
65	66
66	55
72	65
75	57
82	67
88	59
98	75

a. Find the equation of the least-squares line that would allow you to predict the percentage of alumni who would strongly agree that their education was worth the cost, using ranking as the independent variable.

b. Predict the percentage of alumni who would strongly agree that their education was worth the cost for a university with a ranking of 50.

c. Explain why it would not be a good idea to use the least-squares line to predict the percentage of alumni who would strongly agree that their education was worth the cost for a university that had a ranking of 10.

5.20 ● The ability to jump helps salamanders move quickly to avoid predators. The paper **"Jumping in Arboreal Salamanders: A Possible Tradeoff Between Takeoff Velocity and In-Air Posture" (*Zoology* [2020]: 125724)** investigated jumping behavior of 10 species of salamanders and reported the following data on bend angle (the bend [in degrees] in the salamander body, with a larger number reflecting more extreme bending) and the jump takeoff velocity (in meters per second). This data was first introduced in Exercise 5.13, where it was determined that there was a positive correlation between bend angle and takeoff velocity.

Bend Angle	Takeoff Velocity
36.5	0.6
55.0	0.7
43.5	0.7
50.4	0.7
69.8	0.9
71.2	1.0
67.4	1.1
48.1	0.7
61.8	0.9
52.6	0.9

a. If the goal is to learn about how takeoff velocity is related to salamander bend angle, which of these two variables is the response variable and which is the predictor variable?

b. Construct a scatterplot of these data. Would it be reasonable to use a line to summarize the relationship between takeoff velocity and bend angle?

c. Find the equation of the least-squares regression line.

d. Interpret the slope of the least-squares regression line in the context of this study.

e. Does it make sense to interpret the intercept of the least-squares regression line? If so, give an interpretation. If not, explain why it is not appropriate for this data set. (Hint: Think about the range of the x values in the data set.)

f. What would you predict for takeoff velocity for a salamander jumping with a bend angle of 50 degrees?

g. What would you predict for takeoff velocity for a salamander jumping with a bend angle of 65 degrees?

h. Would you use the least-squares regression equation to predict takeoff velocity for a salamander jumping with a bend angle of 90 degrees? If so, what is the predicted takeoff velocity? If not, explain why.

5.21 ● Data on the top speed and maximum height for a random sample of 10 roller coasters in the United States is given in the accompanying table (obtained from the Roller Coaster Data Base, **rcdb.com**).

Top Speed (mph)	Maximum Height (feet)
65	150
55	70
50	161
65	150
60	79
65	157
50	84
52	100
80	209
75	200

a. If the goal is to learn about how top speed is related to maximum height, which of these two variables is the response variable and which is the predictor variable?

b. Construct a scatterplot of the data. How would you describe the relationship between these two variables?

Copyright 2025 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s). Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

c. Find the equation of the least-squares regression line describing the relationship between y = Top speed and x = Maximum height.

d. Interpret the slope of the least-squares regression line in the context of this study.

e. Does it make sense to interpret the intercept of the least-squares regression line? If so, give an interpretation. If not, explain why.

f. What top speed would you predict for a roller coaster with a maximum height of 150 feet?

g. Would you use the least-squares regression line to predict the top speed of a roller coaster with a maximum height of 300 feet? If so, what is the predicted top speed? If not, explain why.

5.22 ● The following table gives data on age, number of cell phone calls made in a typical day, and number of text messages sent in a typical day for a random sample of 10 people selected from those enrolled in adult education classes offered by a school district.

Age	Number of Cell Phone Calls	Number of Text Messages Sent	Age	Number of Cell Phone Calls	Number of Text Messages Sent
55	6	20	19	15	14
33	5	35	30	10	30
60	1	3	33	3	2
38	2	1	37	3	1
55	4	2	52	5	20

An alternative formula for calculating the slope of the least squares regression line that doesn't involve calculating the deviations from the mean for the x values and the y values is

$$b = \frac{\Sigma xy - \dfrac{(\Sigma x)(\Sigma y)}{n}}{\Sigma x^2 - \dfrac{(\Sigma x)^2}{n}}$$

Use this formula to find the equation of the least-squares regression line for predicting y = Number of cell phone calls using x = Age as a predictor.

5.23 The relationship between hospital patient-to-nurse ratio and various characteristics of job satisfaction and patient care has been the focus of a number of research studies. Suppose x = Patient-to-nurse ratio is the independent variable. For each of the following potential dependent variables, indicate whether you expect the slope of the least-squares line to be positive or negative and give a brief explanation for your choice.

a. y = Measure of nurse's job satisfaction (higher values indicate higher satisfaction)

b. y = Measure of patient satisfaction with hospital care (higher values indicate higher satisfaction)

c. y = Measure of patient quality of care

5.24 ● The report **"Airline Quality Rating 2016" (airlinequalityrating.com)** included the accompanying data on the on-time arrival percentage and the number of complaints filed per 100,000 passengers for U.S. airlines.

The report did not include data on the number of complaints for two of the airlines. Use the given data on the other airlines to fit the least-squares line and use it to predict the number of complaints per 100,000 passengers for Spirit and Virgin America Airlines.

Airline	On-Time Arrival Percentage	Complaints per 100,000 Passengers
Alaska	86	0.42
American	80	2.12
Delta	86	0.72
Envoy Air	74	1.59
Express Jet	78	1.01
Frontier	73	3.91
Hawaiian	88	0.89
JetBlue	76	1.17
SkyWest	80	0.84
Southwest	80	0.53
Spirit	69	Not reported
United	78	2.71
Virgin America	80	Not reported

5.25 ● Acrylamide is a chemical that is sometimes found in cooked starchy foods and which is thought to increase the risk of certain kinds of cancer. The paper **"A Statistical Regression Model for the Estimation of Acrylamide Concentrations in French Fries for Excess Lifetime Cancer Risk Assessment" (*Food and Chemical Toxicology* [2012]: 3867–3876)** describes a study to investigate the effect of frying time (in seconds) and acrylamide concentration (in micrograms per kilogram) in French fries. The data in the accompanying table are approximate values read from a graph that appeared in the paper.

Frying Time	Acrylamide Concentration
150	155
240	120
240	190
270	185
300	140
300	270

a. If the goal is to learn how acrylamide concentration is related to frying time, which of these two variables is the dependent variable and which is the independent variable?

b. Construct a scatterplot of these data. Describe any interesting features of the scatterplot.

5.26 Use the acrylamide data given in the previous exercise to answer the following questions.

a. Find the equation of the least-squares line for predicting acrylamide concentration using frying time.

b. Does the equation of the least-squares line support the conclusion that longer frying times tend to be paired with higher acrylamide concentrations? Explain.

c. What is the predicted acrylamide concentration for a frying time of 225 seconds?

d. Would you use the least-squares line to predict acrylamide concentration for a frying time of 500 seconds? If so, what is the predicted concentration? If not, explain why.

5.27 ● Studies have shown that people who suffer sudden cardiac arrest have a better chance of survival if a defibrillator shock is administered very soon after cardiac arrest. How is survival rate related to the time between when cardiac arrest occurs and when the defibrillator shock is delivered? This question is addressed in the paper **"Improving Survival from Sudden Cardiac Arrest: The Role of Home Defibrillators" (by J. K. Stross, University of Michigan, February 2002; available at heartstarthome.com)**.

The accompanying data give y = Survival rate (percent) and x = Mean call-to-shock time (minutes) for a cardiac rehabilitation center (in which cardiac arrests occurred while victims were hospitalized and so the call-to-shock time tended to be short) and for four communities of different sizes:

Mean call-to-shock time, x	2	6	7	9	12
Survival rate, y	90	45	30	5	2

a. Construct a scatterplot for these data. How would you describe the relationship between mean call-to-shock time and survival rate?

b. Find the equation of the least-squares line.

c. Use the least-squares line to predict survival rate for a community with a mean call-to-shock time of 10 minutes.

5.28 The data given in the previous exercise on x = Call-to-shock time (in minutes) and y = Survival rate (percent) were used to compute the equation of the least-squares line, which was

$$\hat{y} = 101.33 - 9.30x$$

The newspaper article **"FDA OKs Use of Home Defibrillators" (*San Luis Obispo Tribune*, November 13, 2002)** reported that "every minute spent waiting for paramedics to arrive with a defibrillator lowers the chance of survival by 10 percent." Is this statement consistent with the given least-squares line? Explain.

5.29 The article **"Analysis of Household Electricity Consumption Using Smart Meter Data" (*International Research Journal of Engineering and Technology* [2020])** used data on x = Daily high temperature (in degrees Celsius) and y = Electricity consumption (in kWH) to fit a least-squares line. The paper states that as temperature increases by one degree Celsius, the average electricity consumption increases by 0.031 kWH. If this statement is correct, what is the slope of the least-squares line. Explain your answer.

5.30 ● The following data on sale price, size, and land-to-building ratio for 10 large industrial properties appeared in the paper **"Using Multiple Regression Analysis in Real Estate Appraisal" (*Appraisal Journal* [2002]: 424–430)**:

Property	Sale Price (millions of dollars)	Size (thousands of sq. ft.)	Land-to-Building Ratio
1	10.6	2,166	2.0
2	2.6	751	3.5
3	30.5	2,422	3.6
4	1.8	224	4.7
5	20.0	3,917	1.7
6	8.0	2,866	2.3
7	10.0	1,698	3.1
8	6.7	1,046	4.8
9	5.8	1,108	7.6
10	4.5	405	17.2

a. Calculate and interpret the value of the correlation coefficient between sale price and size.

b. Calculate and interpret the value of the correlation coefficient between sale price and land-to-building ratio.

c. If you wanted to predict sale price and you could use either size or land-to-building ratio as the basis for making predictions, which would you use? Explain.

d. Based on your choice in Part (c), find the equation of the least-squares line you would use for predicting y = sale price.

5.31 Explain why it can be dangerous to use the least-squares line to obtain predictions for x values that are substantially larger or smaller than those contained in the sample.

5.32 The sales manager of a large company selected a random sample of $n = 10$ salespeople and determined for each one the values of x = Years of sales experience and y = Annual sales (in thousands of dollars). A scatterplot of the resulting (x, y) pairs showed a linear pattern.

a. Suppose that the sample correlation coefficient is $r = 0.75$ and that the average annual sales is $\bar{y} = 100$. If a particular salesperson is 2 standard deviations above the mean in terms of experience, what would you predict for that person's annual sales?

b. If a particular person whose sales experience is 1.5 standard deviations below the average experience is predicted to have an annual sales value that is 1 standard deviation below the average annual sales, what is the value of r?

5.33 Explain why the slope b of the least-squares line always has the same sign (positive or negative) as the sample correlation coefficient r.

5.34 ● The California State Park System Statistical Report for the 2014/2015 Fiscal Year (**parks.ca.gov**) gave the accompanying data on x = Amount of money collected in user fees (in thousands of dollars) and y = Operating cost (in thousands of dollars) for nine state parks in the North Coast Redwoods District.

User Fees (thousands of dollars)	Operating Costs (thousands of dollars)
17	100
74	359
811	3,582
380	1,377
33	241
427	760
500	1,044
734	2,205
760	1,620

a. Construct a scatterplot of these data. Describe any interesting features of the scatterplot.

b. Find the equation of the least-squares line.

c. Is the slope of the least-squares line positive or negative? Is this consistent with your description of the scatterplot in Part (a)?

d. Based on the scatterplot in Part (a), do you think that the value of the correlation coefficient for this data set would be less than 0.5 or greater than 0.5? Explain.

Section 5.3 Assessing the Fit of a Line

Once you have found the equation of the least-squares line, the next step is to consider how effectively the line summarizes the relationship between x and y. Important questions to consider are

1. Is a line an appropriate way to summarize the relationship between the two variables?
2. Are there any unusual aspects of the data set that you need to consider before using the least-squares line to make predictions?
3. If you decide that it is reasonable to use the least-squares line as a basis for prediction, how accurate can you expect predictions based on the line to be?

In this section, graphical and numerical methods are introduced that will allow you to answer these questions.

Most of these methods are based on the vertical deviations of the data points from the least-squares line. These vertical deviations are called residuals, and each represents the difference between an actual y value and the corresponding predicted value, $\hat{y}$, that would result from using the least-squares line to make a prediction.

Predicted Values and Residuals

The predicted value corresponding to the first observation in a data set is obtained by substituting that x value, x_1, into the regression equation to obtain $\hat{y}_1$, where

$$\hat{y}_1 = a + bx_1$$

The difference between the actual y value for the first observation, y_1, and the corresponding predicted value is

$$y_1 - \hat{y}_1$$

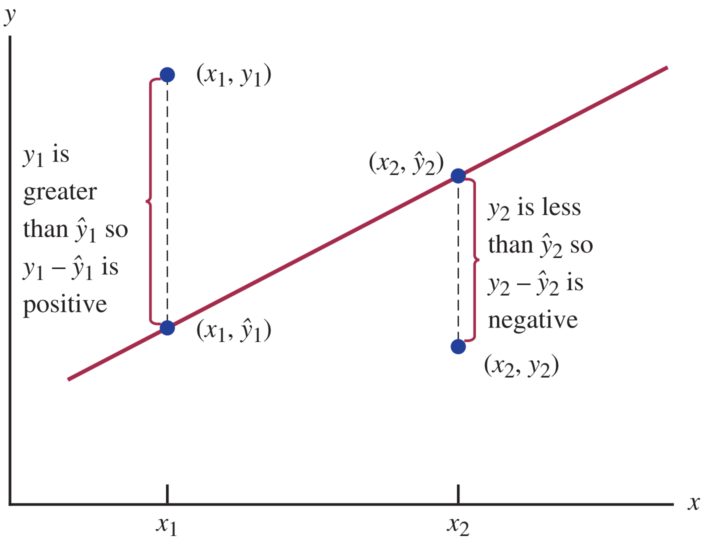
This difference is the **residual**, which is the vertical deviation of a point in the scatterplot from the least-squares line.

Copyright 2025 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s). Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

What Do You Think?

If a residual is positive, what does that tell you about how the actual value compares to the value predicted by the least-squares line? What if the residual is negative?

An observation falling above the line results in a positive residual, whereas a point falling below the line results in a negative residual. This is shown in Figure 5.14.



y
(x_1, y_1)
y_1 is greater than $\hat{y}_1$ so $y_1 - \hat{y}_1$ is positive
$(x_1, \hat{y}_1)$
$(x_2, \hat{y}_2)$
y_2 is less than $\hat{y}_2$ so $y_2 - \hat{y}_2$ is negative
(x_2, y_2)
x_1
x_2
x

Figure 5.14
Positive and negative deviations from the least-squares line (residuals).

Definitions

Predicted values: The **predicted values** result from substituting each sample x value into the equation for the least-squares line. This gives

$$\hat{y}_1 = \textit{first predicted value} = a + bx_1$$

$$\hat{y}_2 = \textit{second predicted value} = a + bx_2$$

$$\vdots$$

$$\hat{y}_n = \textit{nth predicted value} = a + bx_n$$

Residuals: The **residuals** from the least-squares line are the n quantities

$$y_1 - \hat{y}_1 = \textit{first residual}$$

$$y_2 - \hat{y}_2 = \textit{second residual}$$

$$\vdots$$

$$y_n - \hat{y}_n = \textit{nth residual}$$

Each residual is the difference between an observed y value and the corresponding predicted y value.

Example 5.7 Predicting Body Composition and Muscular Strength

Understand the context ➤

● The accompanying data are consistent with information appearing in the paper "Grip Strength Performance as a Determinant of Body Composition, Muscular Strength and Cardiovascular Endurance" (*Journal of Physical Education and Sport* [2022]: 1618–1625). The authors of the paper were interested in whether it would be possible to predict body mass index (BMI) from a simple measure of hand grip strength. The data are given in Table 5.1.

● **Data set available online**

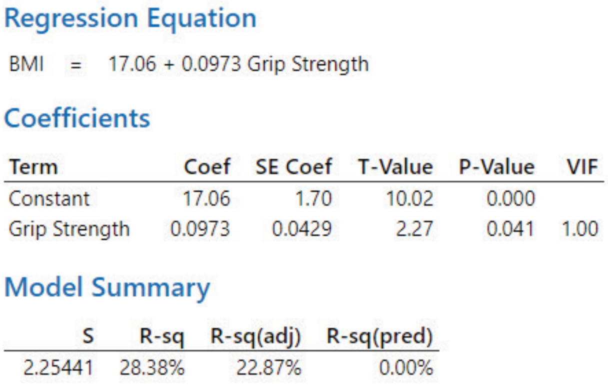
Copyright 2025 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s). Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

Table 5.1 **Predicted Values and Residuals for the Data of Example 5.7**

Consider the data ➤

Grip Strength (x)	BMI (y)	Predicted BMI ($\hat{y}$)	Residual ($y - \hat{y}$)
21.3	19.0	19.13	−0.13
25.2	22.8	19.51	3.29
50.6	22.5	21.99	0.51
17.7	14.0	18.78	−4.78
42.3	22.4	21.18	1.22
32.0	21.7	20.18	1.52
50.4	18.7	21.97	−3.27
35.7	19.8	20.54	−0.74
36.3	20.3	20.59	−0.29
37.0	22.8	20.66	2.14
53.0	21.7	22.22	−0.52
57.6	22.4	22.67	−0.27
57.5	23.4	22.66	0.74
19.2	21.6	18.93	2.67
24.0	17.3	19.40	−2.10

Minitab was used to find the equation of the least-squares line. Partial Minitab output follows:

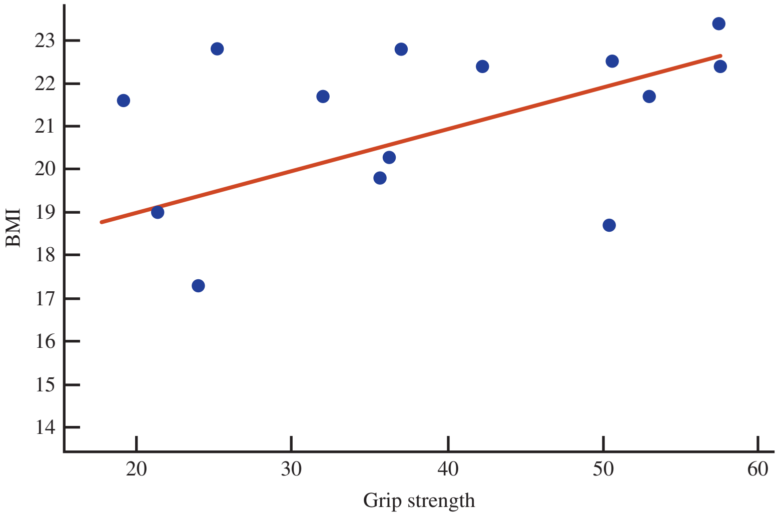


```
Regression Equation

BMI  =  17.06 + 0.0973 Grip Strength

Coefficients

Term             Coef   SE Coef   T-Value   P-Value    VIF
Constant        17.06      1.70     10.02     0.000
Grip Strength  0.0973    0.0429      2.27     0.041   1.00

Model Summary

      S     R-sq   R-sq(adj)   R-sq(pred)
2.25441   28.38%      22.87%        0.00%
```

The resulting least-squares line is $\hat{y} = 17.06 + 0.0973x$.

A scatterplot of the data that also includes the least-squares line is shown in Figure 5.15.

Figure 5.15
Scatterplot for the data of Example 5.7.



Copyright 2025 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s). Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

Do the work ➤ For the person with the smallest *x* value (the fourth observation in the data set, with $x_4 = 17.7$ and $y_4 = 14.0$), the corresponding predicted value and residual are

$$\text{predicted value} = \hat{y}_3 = 17.06 + 0.0973x = 18.78$$

$$\text{residual} = y_4 - \hat{y}_4 = 14.0 - 18.78 = -4.78$$

The other predicted values and residuals are calculated in a similar manner and are included in Table 5.1.

Calculating the predicted values and residuals by hand can be tedious, but Minitab and other statistical software packages, as well as many graphing calculators, include them as part of the output, as shown in Figure 5.16. The predicted values and residuals can be found in the table at the bottom of the Minitab output in the columns labeled “Fit” and “Resid.”

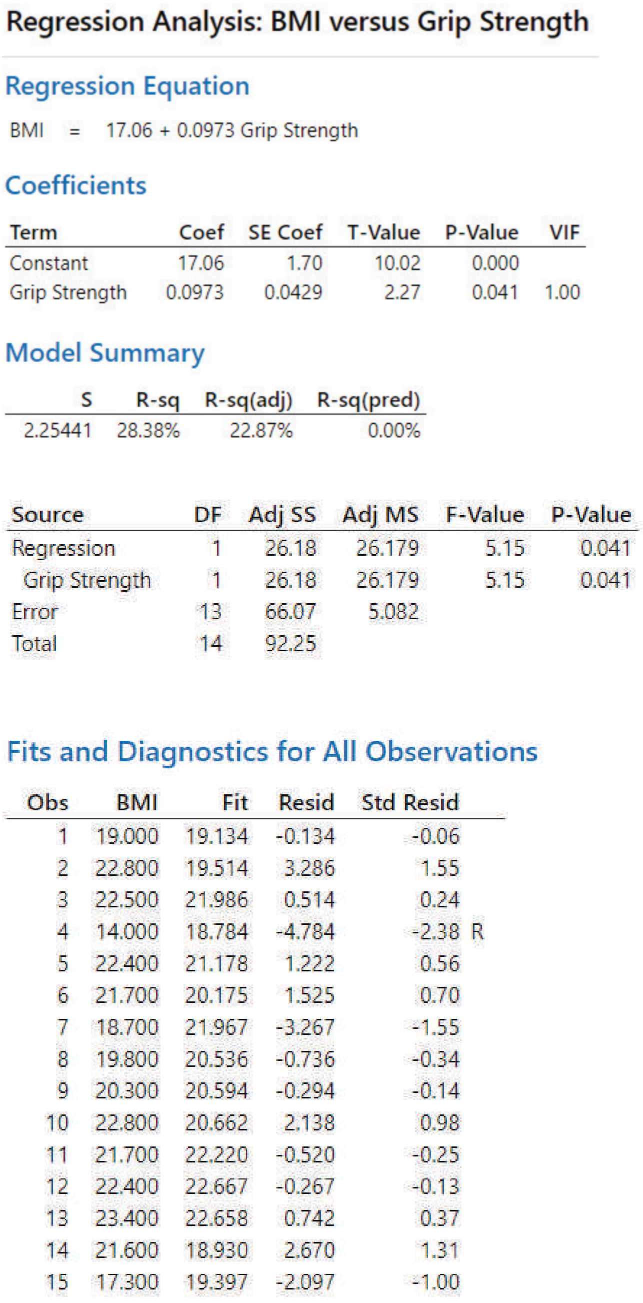


Figure 5.16
Minitab output for the data of Example 5.7.

```
Regression Analysis: BMI versus Grip Strength

Regression Equation

BMI  =  17.06 + 0.0973 Grip Strength

Coefficients

Term            Coef  SE Coef  T-Value  P-Value   VIF
Constant       17.06     1.70    10.02    0.000
Grip Strength 0.0973   0.0429     2.27    0.041  1.00

Model Summary

       S    R-sq  R-sq(adj)  R-sq(pred)
 2.25441  28.38%     22.87%       0.00%


Source           DF  Adj SS  Adj MS  F-Value  P-Value
Regression        1   26.18  26.179     5.15    0.041
  Grip Strength   1   26.18  26.179     5.15    0.041
Error            13   66.07   5.082
Total            14   92.25


Fits and Diagnostics for All Observations

Obs     BMI     Fit   Resid  Std Resid
  1  19.000  19.134  -0.134      -0.06
  2  22.800  19.514   3.286       1.55
  3  22.500  21.986   0.514       0.24
  4  14.000  18.784  -4.784      -2.38  R
  5  22.400  21.178   1.222       0.56
  6  21.700  20.175   1.525       0.70
  7  18.700  21.967  -3.267      -1.55
  8  19.800  20.536  -0.736      -0.34
  9  20.300  20.594  -0.294      -0.14
 10  22.800  20.662   2.138       0.98
 11  21.700  22.220  -0.520      -0.25
 12  22.400  22.667  -0.267      -0.13
 13  23.400  22.658   0.742       0.37
 14  21.600  18.930   2.670       1.31
 15  17.300  19.397  -2.097      -1.00
```

Copyright 2025 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s). Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

Plotting the Residuals

A careful look at residuals can reveal many potential problems. A **residual plot** is a good place to start when assessing the appropriateness of the least-squares line.

Definition

Residual plot: A scatterplot of the (x, residual) pairs.

Isolated points or a pattern of points in a residual plot indicate potential problems.

A desirable residual plot is one that exhibits no particular pattern, such as curvature. Curvature in the residual plot is an indication that the relationship between x and y is not linear and that a curve would be a better choice than a line for describing the relationship between x and y. Curvature is sometimes easier to see in a residual plot than in a scatterplot of y versus x, as illustrated in Example 5.8.

Example 5.8 Record Times

Understand the context ➤

● Consider the accompanying data on x = Distance (in meters) and y = Record time (in seconds) for races of various length in men's international track and field competitions (from *The World Almanac and Book of Facts 2016*).

Distance (in meters)	Record Time (in seconds)
100	10
200	19
400	43
800	101
1000	132
1500	206
2000	285
3000	441
5000	757

The scatterplot, displayed in Figure 5.17(a), appears quite straight. However, even though the value of the correlation coefficient is very close to 1, when the residuals from the least-squares line are plotted (see Figure 5.17(b)), there is a definite curved pattern. Because of this, a curve would be a better choice for describing the relationship between distance and record time.

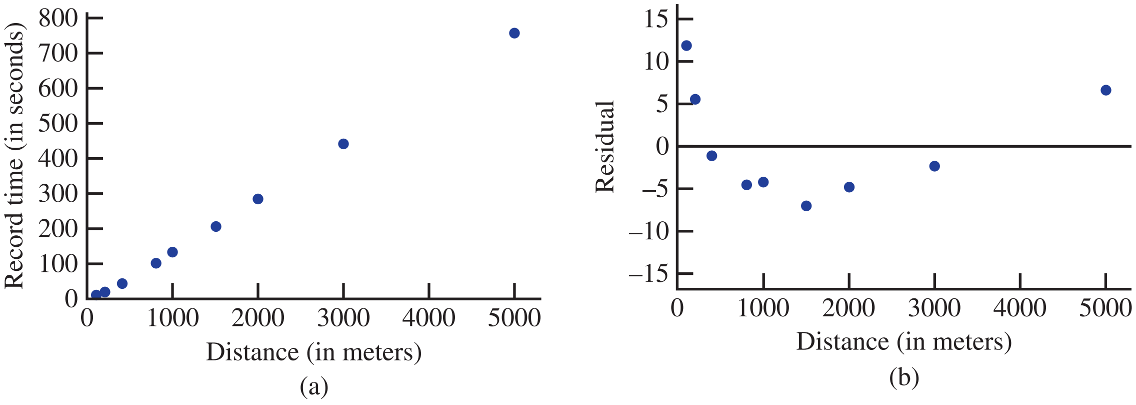


Figure 5.17 Plots for the data of Example 5.8 (a) scatterplot; (b) residual plot.

● **Data set available online**

Copyright 2025 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s). Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

There is another common type of residual plot—one that plots the residuals versus the corresponding $\hat{y}$ values rather than versus the x values. Because $\hat{y} = a + bx$ is simply a linear function of x, the only real difference between the two types of residual plots is the scale on the horizontal axis. The pattern of points in the residual plots will be the same, and it is this pattern of points that is important, not the scale. The two plots give equivalent information, as can be seen in Figure 5.18, which gives both plots for the data of Example 5.7.

It is also important to look for unusual values in the scatterplot or in the residual plot. A point falling far above or below the horizontal line at height 0 corresponds to a large residual, which may indicate some unusual circumstance such as a recording error, a nonstandard experimental condition, or an atypical experimental subject.

A point whose x value differs greatly from others in the data set may have exerted excessive influence in determining the fitted line. One method for assessing the impact of such an isolated point is to delete it from the data set, recompute the equation of the least-squares line, and evaluate the extent to which the equation of the line has changed.

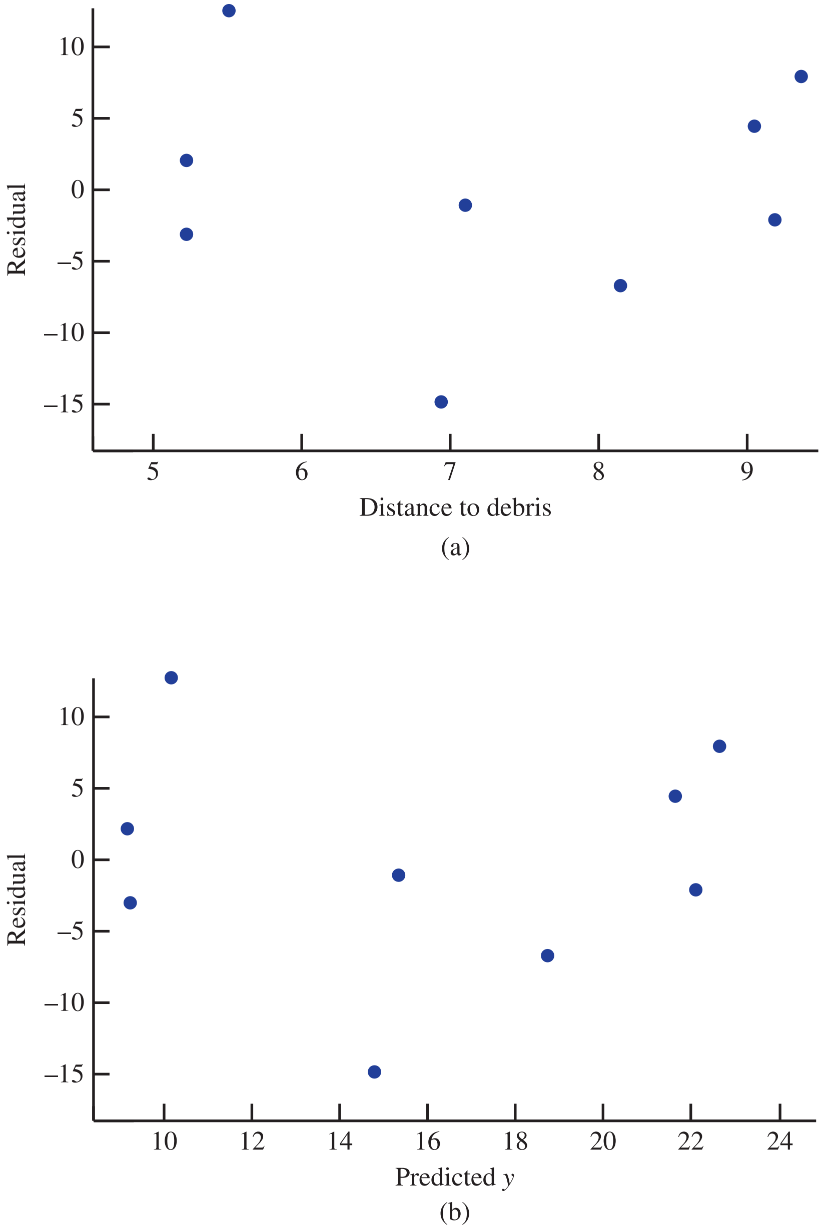


Figure 5.18
Plots for the data of Example 5.7.
(a) plot of residuals versus x;
(b) plot of residuals versus $\hat{y}$.

Copyright 2025 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s). Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

Example 5.9 Older Than Your Average Bear

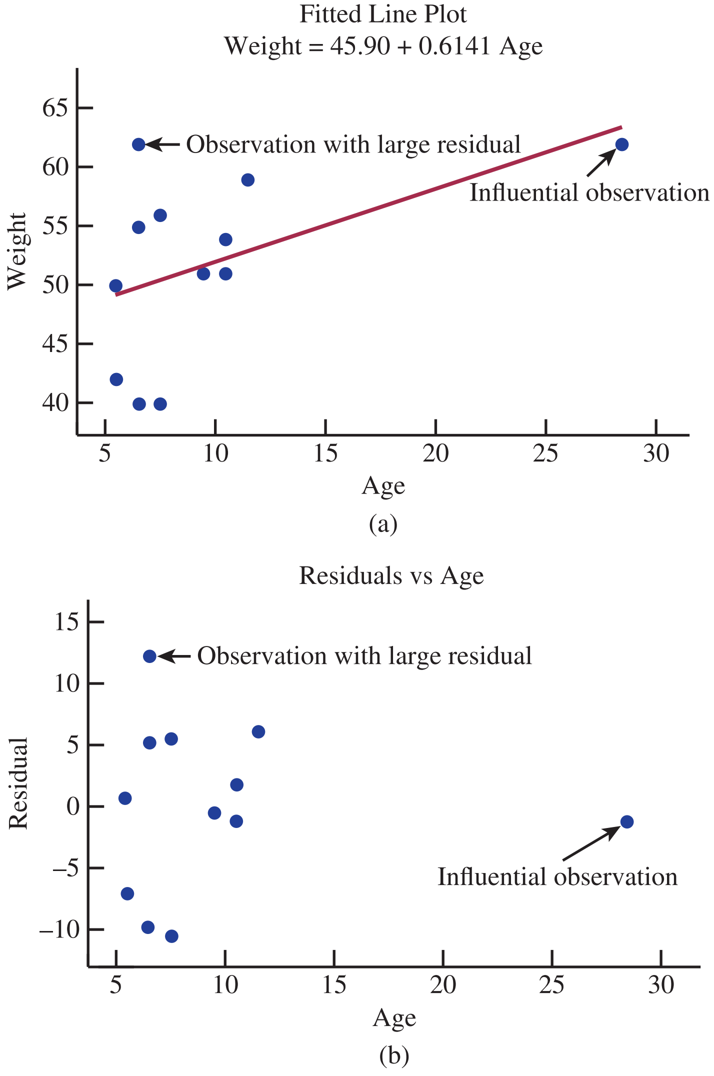
Understand the context ➤

● The accompanying data on x = Age (in years) and y = Weight (in kg) for 12 black bears appeared in the paper **"Habitat Selection by Black Bears in an Intensively Logged Boreal Forest" (*Canadian Journal of Zoology* [2008]: 1307–1316)**.

A scatterplot and residual plot are shown in Figures 5.19(a) and 5.19(b), respectively. One bear in the sample was much older than the other bears (bear 3 with an age of x = 28.5 years and a weight of y = 62.00 kg). This results in a point in the scatterplot that is far to the right of the other points in the scatterplot.

Bear	Age	Weight
1	10.5	54
2	6.5	40
3	28.5	62
4	10.5	51
5	6.5	55
6	7.5	56
7	6.5	62
8	5.5	42
9	7.5	40
10	11.5	59
11	9.5	51
12	5.5	50

Figure 5.19
Plots for the bear data of Example 5.9:
(a) scatterplot;
(b) residual plot.



(a)

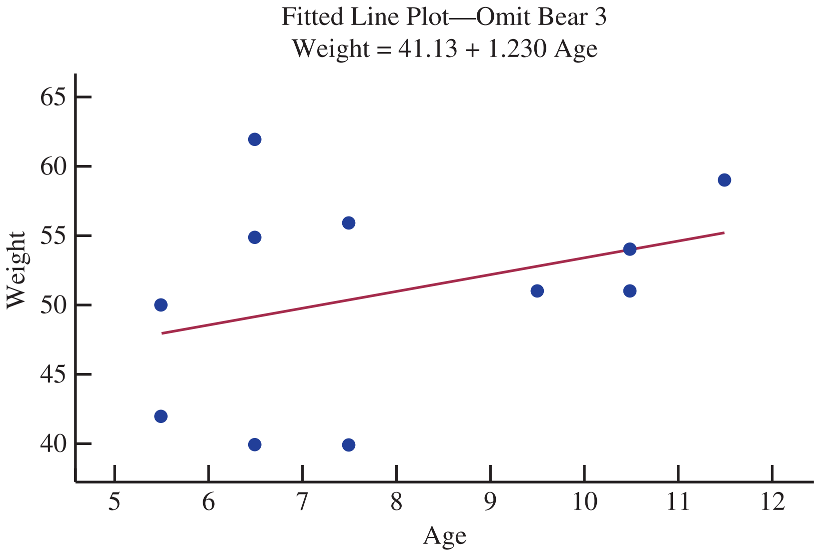
(b)

● **Data set available online**

Copyright 2025 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s). Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

Because the least-squares line minimizes the sum of squared residuals, the line is pulled toward this observation. This single observation plays a big role in determining the slope of the least-squares line, and it is therefore called an **influential observation**. Notice that this influential observation is not necessarily one with a large residual, because the least-squares line actually passes near this point. Figure 5.20 shows what happens when the influential observation is removed from the data set. Both the slope and intercept of the least-squares line are quite different from the slope and intercept of the line when this influential observation is included.

Figure 5.20
Scatterplot and least-squares line with bear 3 removed from data set.



Some points in a scatterplot may fall far from the least-squares line in the y direction, resulting in a large residual. These points are sometimes referred to as **outliers**. In Example 5.9, the observation with the largest residual is bear 7 with an age of $x = 6.5$ years and a weight of $y = 62.00$ kg. This observation is labeled in Figure 5.19. Even though this observation has a large residual, this observation is not very influential. The equation of the least-squares line for the data set consisting of all 12 observations is $\hat{y} = 45.90 + 0.6141x$, which is not much different from the equation that results when bear 7 is deleted from the data set ($\hat{y} = 43.81 + 0.7131x$).

Definitions

Unusual points in a bivariate data set are those that fall away from most of the other points in the scatterplot in either the x direction or the y direction.

Influential observation: An observation is potentially influential if it has an x value that is far away from the rest of the data (separated from the rest of the data in the x direction). To determine if the observation is influential, consider whether removal of this observation has a large impact on the value of the slope or intercept of the least-squares line.

Outlier: An observation that has a large residual. Outlier observations fall far away from the least-squares line in the y direction.

Careful examination of a scatterplot and a residual plot can help you determine the appropriateness of a line for summarizing a relationship. If you decide that a line is appropriate, the next step is to think about assessing the accuracy of predictions based on the least-squares line and deciding whether these predictions (based on the value of x) are better in general than those made without knowledge of the value of x. Two numerical measures that are helpful in this assessment are the *coefficient of determination* and the *standard deviation about the least-squares line*.

Coefficient of Determination, r^2

Suppose that you would like to predict the price of houses in a particular city. A random sample of 20 houses that are for sale is selected, and y = Price and x = Size (in square feet) are recorded for each house in the sample. There will be variability in house price (the houses will differ with respect to price), and it is this variability that makes accurate prediction of price a challenge. How much of the variability in house price can be explained by the fact that price is related to house size and that houses differ in size? If differences in size account for a large proportion of the variability in price, a price prediction that takes house size into account should be a big improvement over a prediction that is not based on size.

The **coefficient of determination** is a measure of the proportion of variability in the y variable that can be "explained" by a linear relationship between x and y.

Definitions

Coefficient of determination: The proportion of variability in y that can be attributed to an approximate linear relationship between x and y. The coefficient of determination is denoted by $\boldsymbol{r^2}$.

The value of r^2 is often converted to a percentage (by multiplying by 100) and interpreted as the percentage of variability in y that can be explained by an approximate linear relationship between x and y.

To understand how r^2 is calculated, first consider variability in the y values. Variability in y can effectively be explained by an approximate straight-line relationship when the points in the scatterplot fall close to the least-squares line—that is, when the residuals are small in magnitude. A natural measure of variability about the least-squares line is the sum of the squared residuals. (Squaring before combining prevents negative and positive residuals from offsetting one another.) A second sum of squares assesses the total amount of variability in observed y values by considering how spread out the y values are from the mean y value.

Definitions

Total sum of squares: The **total sum of squares**, denoted by **SSTo**, is defined as

$$\text{SSTo} = (y_1 - \bar{y})^2 + (y_2 - \bar{y})^2 + \cdots + (y_n - \bar{y})^2 = \Sigma(y - \bar{y})^2$$

Residual sum of squares: The **residual sum of squares** (sometimes referred to as the error sum of squares), denoted by **SSResid**, is defined as

$$\text{SSResid} = (y_1 - \hat{y}_1)^2 + (y_2 - \hat{y}_2)^2 + \cdots + (y_n - \hat{y}_n)^2 = \Sigma(y - \hat{y})^2$$

These sums of squares can be found as part of the regression output from most standard statistical packages or can be obtained using the following formulas:

$$\text{SSTo} = \Sigma y^2 - \frac{\Sigma y^2}{n}$$

$$\text{SSResid} = \Sigma y^2 - a\Sigma y - b\Sigma xy$$

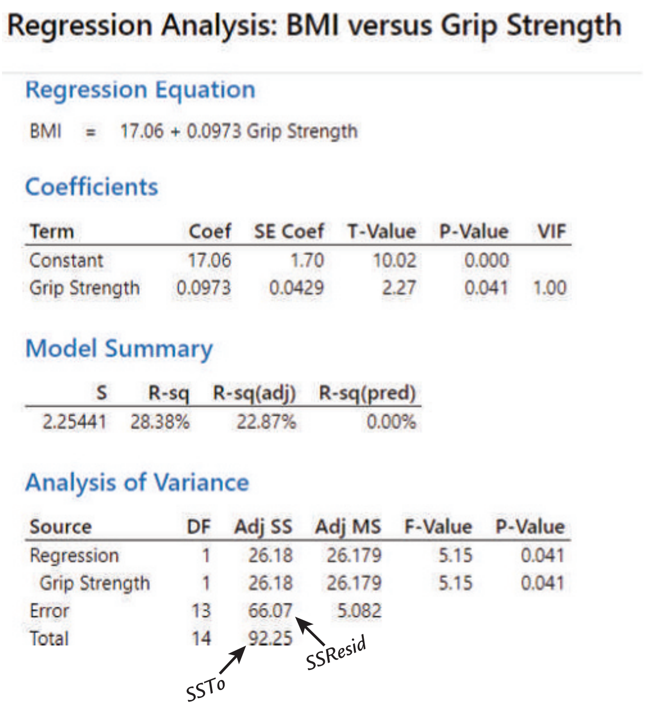
Copyright 2025 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s). Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

Example 5.10 Revisiting the Grip Strength and BMI Data

Figure 5.21 displays part of the Minitab output that results from fitting the least-squares line to the data on y = BMI and x = Grip strength from Example 5.7. From the output,

$$\text{SSTo} = 92.25 \text{ and } \text{SSResid} = 66.07$$

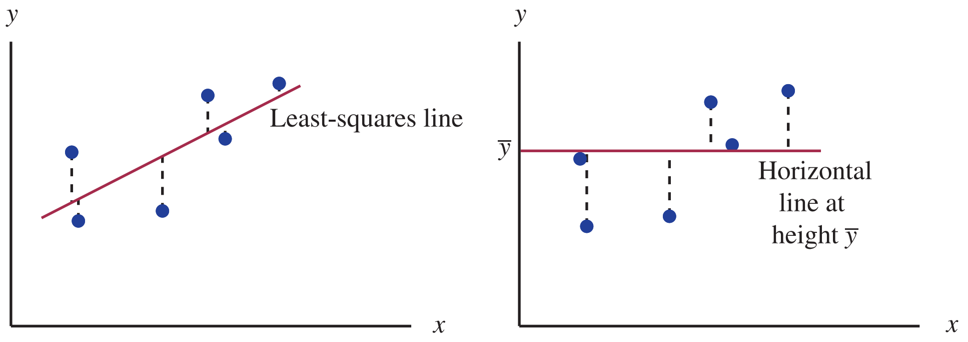
Notice that SSResid is fairly large relative to SSTo.



```
Regression Analysis: BMI versus Grip Strength

Regression Equation

BMI  =  17.06 + 0.0973 Grip Strength

Coefficients

Term              Coef   SE Coef  T-Value  P-Value   VIF
Constant         17.06      1.70    10.02    0.000
Grip Strength   0.0973    0.0429     2.27    0.041   1.00

Model Summary

       S    R-sq   R-sq(adj)   R-sq(pred)
 2.25441  28.38%      22.87%        0.00%

Analysis of Variance

Source           DF  Adj SS  Adj MS  F-Value  P-Value
Regression        1   26.18  26.179     5.15    0.041
  Grip Strength   1   26.18  26.179     5.15    0.041
Error            13   66.07   5.082
Total            14   92.25
```

Figure 5.21
Minitab output for the data of Example 5.10.

The residual sum of squares is the sum of squared vertical deviations from the least-squares line. As Figure 5.22 illustrates, SSTo is also a sum of squared vertical deviations from a line—the horizontal line at height $\bar{y}$. The least-squares line is, by definition, the one having the smallest sum of squared deviations. This means that it is always the case that SSResid $\leq$ SSTo. The two sums of squares are equal only when the least-squares line *is* the horizontal line.



(a) (b)

Figure 5.22
Interpreting sums of squares: (a) SSResid = sum of squared vertical deviations from the least-squares line; (b) SSTo = sum of squared vertical deviations from the horizontal line at $\bar{y}$.

SSResid is sometimes referred to as a measure of unexplained variability—the amount of variability in y that cannot be attributed to the linear relationship between x and y. The more the points in the scatterplot deviate from the least-squares line, the larger the value of SSResid and the greater the amount of variability in y that cannot be explained by the approximate linear relationship. Similarly, SSTo is interpreted as a measure of total variability. The greater the amount of variability in $y_1, y_2, \ldots, y_n$, the greater the value of SSTo.

Copyright 2025 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s). Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

The ratio SSResid/SSTo is the fraction or proportion of total variability that is not explained by a straight-line relationship. Subtracting this ratio from 1 results in the proportion of total variability that *is* explained:

Definition

Coefficient of determination, r^2: The coefficient of determination is defined as

$$r^2 = 1 - \frac{\text{SSResid}}{\text{SSTo}}$$

Multiplying r^2 by 100 gives the percentage of variability in y that can be explained by the approximate linear relationship. The closer this percentage is to 100%, the more successful the linear relationship is in explaining variability in y.

Example 5.11 r^2 for the Grip Strength BMI Data

Do the work ➤ For the data on BMI and grip strength from Example 5.10, SSTo = 92.25 and SSResid = 66.07. It follows that

$$r^2 = 1 - \frac{\text{SSResid}}{\text{SSTo}} = 1 - \frac{66.07}{92.25} = 0.28$$

Interpret the results ➤ This means that only 28% of the observed variability in BMI can be explained by an approximate linear relationship between BMI and grip strength. Notice that the r^2 value can be found in the Minitab output of Figure 5.21, labeled "R-Sq."

The symbol r was used in Section 5.1 to denote the sample correlation coefficient. It is not a coincidence that r^2 is used to represent the coefficient of determination. The notation suggests how these two quantities are related:

$$(\text{correlation coefficient})^2 = \text{coefficient of determination}$$

This means that if $r = 0.8$ or $r = -0.8$, then $r^2 = 0.64$, so 64% of the observed variability in the dependent variable can be explained by the linear relationship. Because the value of r does not depend on which variable is labeled x, the same is true of r^2. The coefficient of determination is one of the few quantities in a regression analysis whose value remains the same when the roles of the dependent and independent variables are interchanged. When $r = 0.5$, $r^2 = 0.25$, so only 25% of the observed variability is explained by a linear relationship. This is why a value of r between -0.5 and 0.5 is not considered evidence of a strong linear relationship. Refer to Figure 5.4 on page 229 for guidance on assessing the strength of a linear relationship.

Example 5.12 Lead Exposure and Brain Volume

Understand the context ➤ The authors of the paper **"Decreased Brain Volume in Adults with Childhood Lead Exposure" (*Public Library of Science Medicine* [May 27, 2008]: e112)** studied the relationship between childhood environmental lead exposure and a measure of brain volume change in a particular region of the brain. Data on x = Mean childhood blood lead level (μg/dL) and y = Brain volume change (percent) read from a graph that appeared in the paper were used to produce the scatterplot in Figure 5.23. The least-squares line is also shown on the scatterplot.

Interpret the results ➤ Figure 5.24 displays computer output that results from fitting the least-squares line to the data. Notice that although there is a slight tendency for smaller y values (corresponding to a brain volume decrease) to be paired with higher values of mean blood lead levels, the relationship is weak. The points in the plot are widely scattered around the least-squares line.

Copyright 2025 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s). Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

Figure 5.23
Scatterplot and least-squares line for the data of Example 5.12.

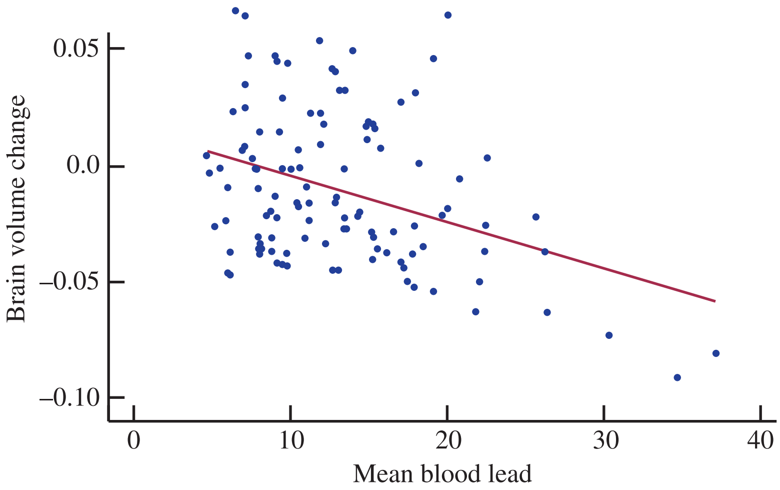


Figure 5.24
Regression output for the data of Example 5.12.

```
Regression Analysis: Brain Volume Change versus Mean Blood Lead

The regression equation is
Brain Volume Change = 0.01559 – 0.001993 Mean Blood Lead

S = 0.0310931      R-Sq = 13.6%      R-Sq(adj) = 12.9%

Analysis of Variance

Source        DF        SS          MS         F       P
Regression     1    0.016941   0.0169410   17.52   0.000
Error        111    0.107313   0.0009668
Total        112    0.124254
```

From the computer output, you can see that $100r^2 = 13.6\%$, so $r^2 = 0.136$. This means that differences in childhood mean blood lead level explain only 13.6% of the variability in adult brain volume change. Because the coefficient of determination is the square of the correlation coefficient, you can calculate the value of the correlation coefficient by taking the square root of r^2. In this case, you know that the correlation coefficient will be negative (because there is a negative relationship between x and y), so you want the negative square root:

$$r = -\sqrt{0.136} = -0.369$$

Based on the values of the correlation coefficient and the coefficient of determination, you would conclude that there is a weak negative linear relationship and that childhood mean blood lead level explains only about 13.6% of the variability in adult change in brain volume.

Standard Deviation About the Least-Squares Line, s_e

The coefficient of determination measures the variability about the least-squares line *relative* to overall variability in y. A large value of r^2 does not by itself promise that the deviations from the line are small in an absolute sense. A typical observation could deviate from the line by quite a bit, but these deviations might still be small relative to overall variability in y.

Recall that in Chapter 4 the sample standard deviation

$$s = \sqrt{\frac{\Sigma(x - \bar{x})^2}{n - 1}}$$

was used as a measure of variability in a single sample. Roughly speaking, s is a typical amount that a sample observation deviates from the sample mean. There is a similar measure of variability that is used when a least-squares line is fit. This measure of variability is the standard deviation about the least-squares line.

Definition

Standard deviation about the least-squares line: The standard deviation about the least-squares line is defined as

$$s_e = \sqrt{\frac{\text{SSResid}}{n-2}}$$

From a practical point of view, s_e can be interpreted as a typical amount that an observation deviates from the least-squares line.

The standard deviation about the least-squares line is given as part of the usual regression output when statistical software or a graphing calculator is used. For the brain volume data of Example 5.12, the value of s_e can be found in the Minitab output of Figure 5.24 labeled as "S." For the brain volume data, $s_e \approx 0.031$.

Example 5.13 Predicting Graduation Rates

Understand the context ➤

● Consider the accompanying data from 2021 on six-year graduation rate (%), instructional expenditure per full-time student (in dollars), and student-faculty ratio (the number of students per faculty member) for 11 primarily undergraduate public universities and colleges in the western United States with enrollments between 10,000 and 20,000 **(Source: National Center for Education Statistics, nces.ed.gov)**.

Consider the data ➤

Graduation Rate	Instructional Expenditure	Student-Faculty Ratio
43	5,078	21
67	6,040	22
44	4,766	25
50	4,770	24
55	5,918	25
54	5,570	25
59	5,979	20
57	7,762	15
54	5,368	20
75	11,581	24
68	9,388	17

Figure 5.25 displays scatterplots of graduation rate versus instructional expenditure and graduation rate versus student-faculty ratio. The least-squares lines and the values of r^2 and s_e are also shown.

Interpret the results ➤

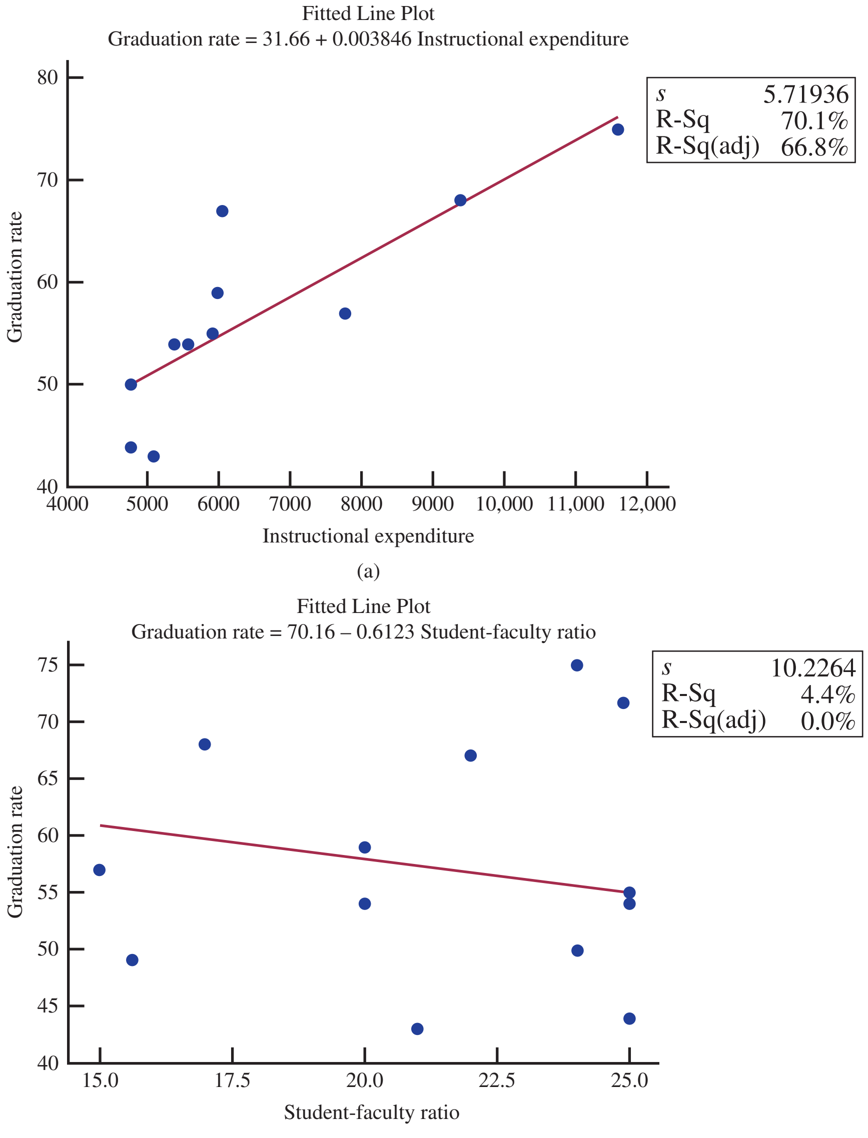
Notice that there is a positive linear relationship between instructional expenditure and graduation rate, but there does not appear to be much of a relationship between student-faculty ratio and graduation rate. For instructional expenditure, the value of r^2 is 0.701 (70.1%), indicating that about 70.1% of the variability in graduation rate from university to university can be explained by differences in instructional expenditures. The standard deviation about the least-squares line is $s_e = 5.72$, compared to 10.23 for student-faculty ratio. This means that the points in the scatterplot of graduation rate versus instructional expenditure tend to fall closer to the least-squares line than is the case for the line that describes graduation rate versus student-faculty ratio.

The value of r^2 for graduation rate versus instructional expense is greater than the value of r^2 for graduation rate versus student-faculty ratio and the value of s_e is smaller than for graduation rate versus student-faculty ratio, so instructional expense does a better job of explaining variability in graduation rates. The least-squares line that uses educational

● **Data set available online**

expense as a predictor would produce more accurate estimates of graduation rates than the line that uses student-faculty ratio as a predictor.

Figure 5.25
Scatterplots for the data of Example 5.13:
(a) graduation rate versus instructional expenditure;
(b) graduation rate versus student-faculty ratio.



(a)

(b)

Also, take a second look at the scatterplot of Graduation rate versus Instructional expenditures. There are two schools that stand out in the scatterplot as potentially influential observations. The two points in the upper right-hand corner of the plot are far removed from the others in the x direction, indicating that these two universities had noticeably higher expenditures than the other nine universities. If these two data points are removed, the equation of the least-squares line changes dramatically. For this smaller data set, the equation of the least-squares line is

$$\widehat{\text{Graduation rate}} = 26.56 + 0.005 \text{ Instructional expenditures}$$

and r^2 is only 34.74%.

What Do You Think?

When using a least-squares regression line to make a prediction, is it better to have a large value of s_e or a small value? Why?

When evaluating the usefulness of the least-squares line for making predictions, it is important to consider both the value of r^2 *and* the value of s_e. These two measures assess different aspects of the fit of the line. In general, you would like to have a small value for s_e (which indicates that deviations from the line tend to be small) and a large value for r^2 (which indicates that the linear relationship explains a large proportion of the variability in the y values).

Interpreting the Values of s_e and r^2

A small value of s_e indicates that residuals tend to be small. Because a residual represents the difference between a predicted y value and an observed y value, the value of s_e tells you about the accuracy you can expect when using the least-squares line to make predictions.

A large value of r^2 indicates that a large proportion of the variability in y can be explained by the approximate linear relationship between x and y. This tells you that knowing the value of x is helpful for predicting y.

A useful least-squares line will have a reasonably small value of s_e and a reasonably large value of r^2.

Now that all of the parts of a linear regression analysis have been considered, let's put them together. The steps in a linear regression analysis are summarized in the accompanying box.

Steps in a Linear Regression Analysis

Given a bivariate numerical data set consisting of observations on a dependent variable y and an independent variable x:

Step 1. Summarize the data graphically by constructing a scatterplot.
Step 2. Based on the scatterplot, decide if it looks like the relationship between x and y is approximately linear. If so, proceed to the next step.
Step 3. Find the equation of the least-squares line.
Step 4. Construct a residual plot and look for any patterns or unusual features that may indicate that a line is not the best way to summarize the relationship between x and y. If none are found, proceed to the next step.
Step 5. Compute the values of s_e and r^2 and interpret them in context.
Step 6. Based on what you have learned from the residual plot and the values of s_e and r^2, decide whether the least-squares line is useful for making predictions. If so, proceed to the last step.
Step 7. Use the least-squares line to make predictions.

Let's return to the example from the chapter introduction and look at how following these steps can help you learn about how the weight of an elephant seal might be estimated.

Example 5.14 Revisiting Elephant Seals

Understand the context ➤

Elephant seals are very large marine mammals that spend most of their time in the ocean. However, each winter, they leave the water and spend about a month on the beach to mate and for the females to give birth to the pups that they have been carrying for a year. Body weight is one of the variables that biologists study in order to monitor the health of elephant seals.

It isn't easy to actually weigh an elephant seal (they are really big!). Because elephant seals are so difficult to weigh, researchers have considered ways to estimate weight so that they don't have to actually weigh them. In a study described in the

Copyright 2025 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s). Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

paper **"Calibration of Aerial Photogrammetry to Estimate Elephant Seal Mass" (*Marine Mammal Science* [2020])**, researchers weighed 44 adult female elephant seals and also took pictures of the seals. They used the pictures to calculate a measure of size that they called the "footprint" of the seal. By studying the relationship between weight and footprint determined from a picture, they hoped to be able to estimate the weight of an elephant seal from its footprint.

Because the researchers were interested in estimating the weight using the measurement of footprint, the dependent variable is x = Weight and the independent variable is footprint. Weight was measured in kg and the footprint calculated from pictures was measured in m^2. The data used in this example are approximate values read from a graph that appeared in the paper.

Step 1: Data from the 44 adult female elephant seals was used to construct a scatterplot, which is shown in Figure 5.26.

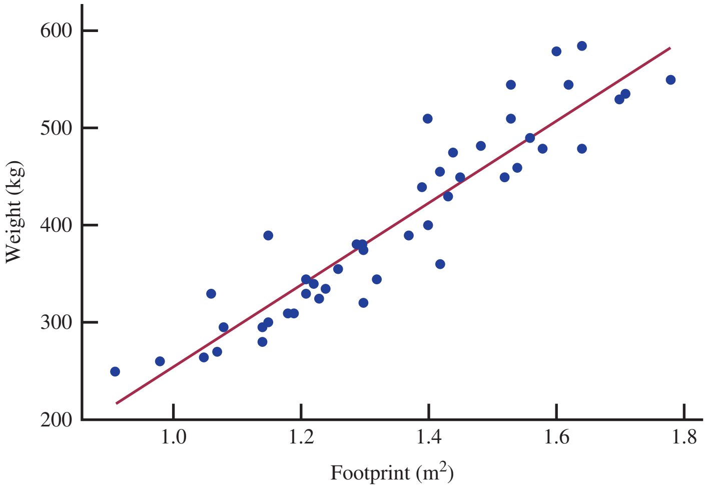
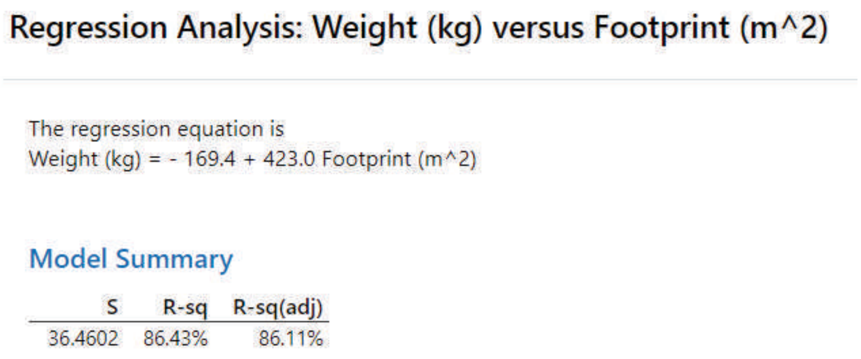


Figure 5.26
Scatterplot of weight versus footprint.

Step 2: Based on the scatterplot, it does appear that there is a reasonably strong positive linear relationship between weight and footprint. The correlation coefficient for these data is $r = 0.93$, which is consistent with the strong positive linear pattern in the scatterplot.

Step 3: Minitab regression output is shown below. The equation of the least-squares line is

$$\hat{y} = -169.4 + 423.0x$$



```
Regression Analysis: Weight (kg) versus Footprint (m^2)

The regression equation is
Weight (kg) = - 169.4 + 423.0 Footprint (m^2)

Model Summary

      S      R-sq   R-sq(adj)
36.4602    86.43%      86.11%
```

Step 4: A residual plot constructed from these data shows a few observations with large residuals, but these observations were not far removed from the rest of the data in the x direction. These observations were not judged to be influential. Also, there were no unusual

Copyright 2025 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s). Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

patterns in the residual plot that would suggest a nonlinear relationship between weight and footprint.

Step 5: The values of s_e and r^2 were

$$s_e = 36.5 \text{ and } r^2 = 0.864$$

This means that approximately 86.4% of the variability in weight can be explained by the linear relationship between weight and footprint. The value of s_e tells you that if the least-squares line is used to predict weight using footprint, a typical difference between the predicted weight and the actual weight would be about 36.5 kg.

Step 6: Based on the residual plot, the large value of r^2, and the relatively small value of s_e, it is reasonable to use footprint and the least-squares line as a way to estimate the weights of adult female elephant seals.

Step 7: To illustrate predicting weight, suppose that from pictures taken of an adult female elephant seal, the seal's footprint is calculated to be 1.5 m^2. The predicted weight of the seal would be

$$\begin{aligned}\hat{y} &= -169.4 + 423.0(1.5)\\ &= -169.4 + 634.5\\ &= 465.1 \text{ kg}\end{aligned}$$

This is just an estimate of the actual weight and a typical prediction error is about 36.5 kg.

Exercises 5.35–5.53

● **Data set available online**

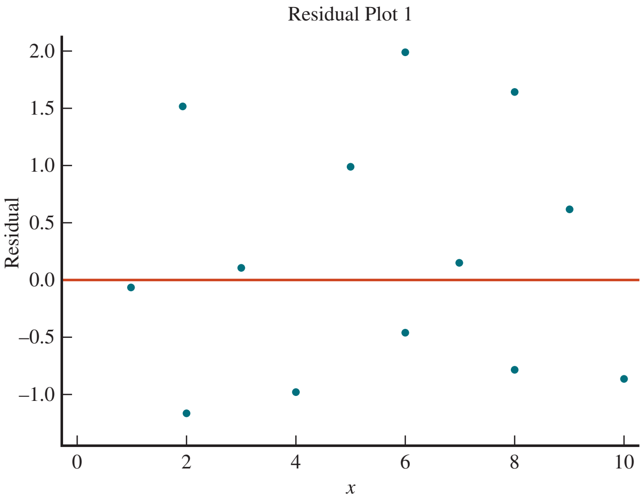
5.35 ● Does it pay to stay in school? The report ***Trends in Higher Education* (The College Board, 2010)** looked at the median hourly wage gain per additional year of schooling. The report states that workers with a high school diploma had a median hourly wage that was 10% higher than those who had only completed 11 years of school. Workers who had completed 1 year of college (13 years of education) had a median hourly wage that was 11% higher than that of the workers who had completed only 12 years of school. The added gain in median hourly wage for each additional year of school is shown in the accompanying table. The entry for 15 years of schooling has been intentionally omitted from the table.

Years of Schooling	Median Hourly Wage Gain for the Additional Year (percent)
12	10
13	11
14	13
16	16
17	18
18	19

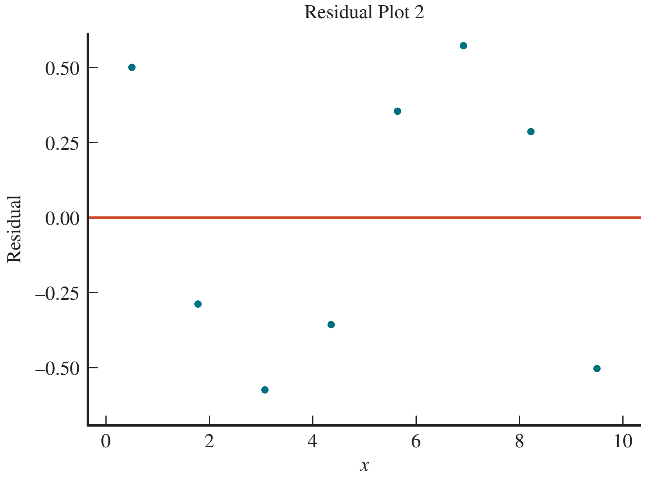
a. Use the given data to predict the median hourly wage gain for the 15th year of schooling.

b. The actual wage gain for 15th year of schooling was 14%. How close was the actual value to the predicted wage gain percent from Part (a)?

5.36 For each of the two residual plots shown, explain whether the least squares regression line would be the most appropriate way to model the relationship between x and y.



Copyright 2025 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s). Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.



5.37 ● The paper **"Unhappiness and Age" (*Journal of Economic Behavior and Organization* [2020]: 461–488)** explores the relationships between age and several measures of mental health and well-being. Based on a large sample of people in the United States and the European Union, the data in the accompanying table are the percentages reporting that they get less than 7 hours of sleep per night by age group.

Age Group	Representative Age	Percent Reporting Less Than 7 Hours of Sleep
20 to 29	25	6
30 to 39	35	9
40 to 49	45	9
50 to 59	55	8
60 to 69	65	1

a. Find the equation of the least-squares regression line that describes the relationship y = Percent reporting less than 7 hours of sleep and x = Representative age.

b. Calculate the five residuals and construct a residual plot.

c. Based on the residual plot, would you recommend using the least squares regression line to describe the relationship between percent reporting less than 7 hours of sleep and age, or would you recommend using a curve rather than a line? What aspect of the residual plot supports your decision?

5.38 ● The paper referenced in the previous exercise also gave the data in the accompanying table on average stress levels and percent reporting less than 7 hours of sleep for people in the five age groups. Higher values for stress level indicate a greater level of stress reported for the people in that age group.

Age Group	Average Stress Level	Percent Reporting Less Than 7 Hours of Sleep
20 to 29	32	6
30 to 39	35	9
40 to 49	36	9
50 to 59	34	8
60 to 69	28	1

a. Construct a scatterplot of y = Average stress level versus x = Percent reporting less than 7 hours of sleep.

b. There is one observation that stands out in the scatterplot as being potentially influential because it has an x value that is distant from the other observations. Which age group corresponds to this observation?

c. Based just on looking at the scatterplot, do you think that observation is influential? Explain your thinking, referring to the scatterplot.

d. Decide if the observation you identified is actually influential by determining if omitting this value results in substantial changes to the slope or intercept of the least-squares regression line. Does this support your answer to Part (c)?

5.39 ● The report **"Airline Quality Rating 2021"** included the data for 10 U.S. airlines given in the table below.

Airline	On-Time Arrival Percentage	Airline Quality Rating
Alaska	85	4
Allegiant	71	2
American	83	7
Delta	86	3
Frontier	84	10
Hawaiian	88	9
JetBlue	82	6
Southwest	86	1
Spirit	87	5
United	84	8

a. With x = Airline quality rating and y = On-time arrival percentage, construct a scatterplot. Does the pattern in the scatterplot look linear?

b. Find the equation of the least-squares regression line.

c. Calculate the residuals and construct a residual plot. Are there any unusual features in the residual plot?

5.40 ● Acrylamide is a chemical that is sometimes found in cooked starchy foods and which is thought to increase the risk of certain kinds of cancer. The paper **"A Statistical Regression Model for the Estimation of Acrylamide Concentrations in French Fries for Excess Lifetime Cancer Risk Assessment" (*Food*

***and Chemical Toxicology* [2012]: 3867–3876)** describes a study to investigate the effect of x = Frying time (in seconds) and y = Acrylamide concentration (in micrograms per kilogram) in French fries. The data in the accompanying table are approximate values read from a graph that appeared in the paper.

Frying Time	Acrylamide Concentration
150	155
240	120
240	190
270	185
300	140
300	270

a. Construct a scatterplot of these data.
b. Find the equation of the least-squares line. Based on this line, what would you predict acrylamide concentration to be for a frying time of 270 seconds? What is the residual associated with the observation (270, 185)?

5.41 Consider the scatterplot of acrylamide concentration versus frying time from the previous exercise.
a. Which observation is potentially influential? Explain the reason for your choice.
b. When the potentially influential observation is deleted from the data set, the equation of the least-squares line using the remaining five observations is $\hat{y} = -44 + 0.83x$. Use this equation to predict the acrylamide concentration for a frying time of 270 seconds. How does this prediction compare to the prediction made in the previous exercise?

5.42 ● Some types of algae have the potential to cause damage to river ecosystems. The accompanying data on algae colony density (y) and rock surface area (x) for nine rivers is a subset of data that appeared in a scatterplot in a paper in the journal ***Aquatic Ecology* (2010: 33–40)**.

x	50	55	50	79	44	37	70	45	49
y	152	48	22	35	38	171	13	185	25

a. Find the equation of the least-squares line.
b. What is the value of r^2 for this data set? Write a sentence interpreting this value in context. (Hint: See Example 5.12.)
c. What is the value of s_e for this data set? Write a sentence interpreting this value in context. (Hint: See Example 5.13.)
d. Is the linear relationship between rock surface area and algae colony density positive or negative? Is it weak, moderate, or strong? Justify your answer.

5.43 ● The paper **"Unhappiness and Age" (*Journal of Economic Behavior and Organization* [2020]: 461–488)** explores the relationships between age and several measures of mental health and well-being. Based on a large sample of people in the United States and the European Union, the data in the accompanying table are the average stress level reported by age group. Higher values for stress level indicate a greater level of stress reported for people in that age group.

Age Group	Representative Age	Average Stress Level
20 to 29	25	32
30 to 39	35	35
40 to 49	45	36
50 to 59	55	34
60 to 69	65	28

a. Find the equation of the least-squares regression line that describes the relationship y = Average stress level and x = Representative age.
b. Calculate the five residuals and construct a residual plot.
c. Based on the residual plot, would you recommend using the least-squares regression line to describe the relationship between stress level and age, or would you recommend using a curve rather than a line? What aspect of the residual plot supports your decision?

5.44 ● The paper referenced in the previous exercise also gave the data in the accompanying table on average stress levels and average anxiety levels for people in the five age groups. Higher values for stress level and anxiety level indicate a greater level of stress or anxiety reported for the people in that age group.

Age Group	Average Stress Level	Average Anxiety Level
20 to 29	32	2.96
30 to 39	35	2.97
40 to 49	36	3.05
50 to 59	34	3.04
60 to 69	28	2.74

a. Construct a scatterplot of y = Average anxiety level versus x = Average stress level.
b. There is one observation that stands out in the scatterplot as being potentially influential because it has an x value that is distant from the

Copyright 2025 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s). Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

other observations. Which age group corresponds to this observation?

c. Based just on looking at the scatterplot, do you think that that observation is influential? Explain your thinking, referring to the scatterplot.

d. Decide if the observation you identified is actually influential by determining if omitting this value results in substantial changes to the slope or intercept of the least-squares regression line. Does this support your answer to Part (c)?

5.45 ● The first Batman movie was made over 50 years ago in 1966. Over the years, Batman has been played on screen by a number of actors and even by a Lego figure in the Lego Batman movies. In the original comic books, Batman was 188 centimeters tall (about 6'2") and weighed 95 kilograms (about 210 pounds). The article **"50 Years of Batman on Film: How Has His Physique Changed?" (economist.com, March 28, 2016)** included the heights and weights of all the onscreen Batmen in the table below.

Batman	Height (cm)	Weight (kg)
Comic book	188	95
Lego Batman	4	4
Adam West	188	91
Michael Keaton	178	72
Val Kilmer	183	93
George Clooney	180	78
Christian Bale	183	82
Ben Affleck	193	98

With x = Height and y = Weight, the equation of the least-squares line is $\hat{y} = 1.052 + 0.466x$.

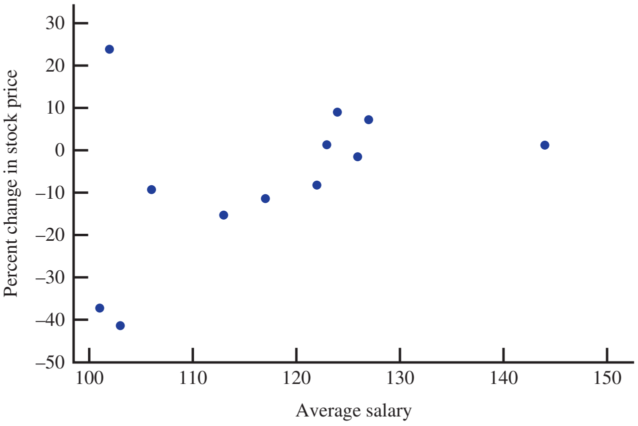
a. Calculate the residuals.

b. Construct a residual plot. Are there any unusual features of the residual plot?

c. The observation for Lego Batman (4, 4) is far removed from the other values in the data set (not surprising!). Is this observation influential in determining either the slope of the least-squares line, or the intercept, or both? Justify your answer.

5.46 ● The accompanying table gives the average yearly salary (in thousands of dollars) in 2022 and the percent change in stock price for the 2022 year for the 12 highest-paying publicly traded companies in the United States. (Data on average salary is from 247wallst.com and percent change in stock prices were calculated from data on **finance.yahoo.com**). A scatterplot of these data is also shown.

Company	Average Salary	Percent Change in Stock Price
Credit Karma	144	1.3
Meta	127	7.2
VMWare	126	−1.6
Palo Alto Networks	124	8.9
Microsoft	123	1.3
Alphabet	122	−8.2
Adobe	117	−11.3
UiPath	113	−15.3
IBM	106	−9.3
Zoom Video	103	−41.3
Calix	102	23.8
Peloton	101	−37.1



The equation of the least-squares line with y = Percent change in stock price and x = Average salary is $\hat{y} = -79.0 + 0.615x$.

a. Identify the observation with the largest residual.

b. The observation for Credit Karma is far removed from the other points in the scatterplot in the x direction. Is this an influential observation? What impact does omitting the observation for Credit Karma from the data set have on the equation of the least-squares line?

c. Explain the difference between an influential observation and an outlier in a bivariate data set.

5.47 The article **"Examined Life: What Stanley H. Kaplan Taught Us About the SAT" (*The New Yorker* [December 17, 2001]: 86–92)** included a summary of findings regarding the use of SAT I scores, SAT II scores, and high school grade point average (GPA) to predict first-year college GPA. The article states that “among these, SAT II scores are the best predictor, explaining 16 percent of the variance in first-year college grades. GPA was second at 15.4 percent, and SAT I was last at 13.3 percent.”

a. If the data from this study were used to fit the least-squares line with y = First-year college GPA and x = High school GPA, what would be the value of r^2?

b. The article stated that SAT II was the best predictor of first-year college grades. Do you think that predictions based on a least-squares line with y = First-year college GPA and x = SAT II score would be very accurate? Explain why or why not.

5.48 ● The accompanying data are a subset of data from the report **“Great Jobs, Great Lives” (Gallup-Purdue Index 2015 Report)**. The values are approximate values read from a scatterplot. Students at a number of universities were asked if they agreed that their education was worth the cost. One variable in the table is the percentage of students at the university who responded *strongly agree*. The other variable in the table is the *U.S. News and World Report* ranking of the university.

Ranking	Percentage of Alumni Who Strongly Agree
28	53
29	58
30	62
37	55
45	54
47	62
52	55
54	62
57	70
60	58
65	66
66	55
72	65
75	57
82	67
88	59
98	75

a. What is the value of r^2 for this data set? Write a sentence interpreting this value in context.

b. What is the value of s_e for this data set? Write a sentence interpreting this value in context.

c. Is the linear relationship between percentage of alumni who think their education was worth the cost and university ranking positive or negative? Is it weak, moderate, or strong? Justify your answer.

5.49 ● The article **“California State Parks Closure List Due Soon” (*The Sacramento Bee,* August 30, 2009)** gave the following data on y = Number of employees in fiscal year 2007–2008 and x = Total size of parks (in acres) for the 20 state park districts in California:

Number of Employees, y	Total Park Size, x
95	39,334
95	324
102	17,315
69	8,244
67	620,231
77	43,501
81	8,625
116	31,572
51	14,276
36	21,094
96	103,289
71	130,023
76	16,068
112	3,286
43	24,089
87	6,309
131	14,502
138	62,595
80	23,666
52	35,833

a. Construct a scatterplot of the data.

b. Find the equation of the least-squares line.

c. Do you think the least-squares line gives accurate predictions? Explain.

d. Delete the observation with the largest x value from the data set and recalculate the equation of the least-squares line. Does this observation greatly affect the equation of the line?

5.50 ● The article referenced in the previous exercise also gave data on the percentage of operating costs covered by park revenues for the 2007–2008 fiscal year.

Number of Employees, x	Percent of Operating Cost Covered by Park Revenues, y
95	37
95	19
102	32
69	80
67	17
77	34
81	36
116	32

(continued)

Copyright 2025 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s). Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

Number of Employees, *x*	Percent of Operating Cost Covered by Park Revenues, *y*
51	38
36	40
96	53
71	31
76	35
112	108
43	34
87	97
131	62
138	36
80	36
52	34

a. Find the equation of the least-squares line relating y = Percent of operating costs covered by park revenues and x = Number of employees.

b. Based on the values of r^2 and s_e, do you think that the least-squares line does a good job of describing the relationship between y = Percent of operating costs covered by park revenues and x = Number of employees? Explain.

c. The graph at the bottom of the page is a scatterplot of y = Percent of operating costs covered by park revenues and x = Number of employees. The least-squares line is also shown. Which observations are outliers? Do the observations with the largest residuals correspond to the park districts with the largest number of employees?

5.51 A study was carried out to investigate the relationship between the hardness of molded plastic (y, in Brinell units) and the amount of time elapsed since the plastic was molded (x, in hours). Summary quantities include $n = 15$, SSResid = 1235.470, and SSTo = 25,321.368. Calculate and interpret the value of the coefficient of determination.

5.52 Both r^2 and s_e are used to assess the fit of a line.

a. Is it possible that both r^2 and s_e could be large for a bivariate data set? Explain. (A picture might be helpful.)

b. Is it possible that a bivariate data set could yield values of r^2 and s_e that are both small? Explain. (Again, a picture might be helpful.)

c. Explain why it is desirable to have r^2 large and s_e small if the relationship between two variables x and y is to be described using a straight line.

5.53 With a bit of algebra, it can be shown that

$$\text{SSResid} = (1 - r^2)\Sigma(y - \bar{y})^2$$

from which it follows that

$$s_e = \sqrt{\frac{n-1}{n-2}}\sqrt{1 - r^2}\, s_y$$

Unless n is quite small, $(n - 1)/(n - 2) \approx 1$, so

$$s_e \approx \sqrt{1 - r^2}\, s_y$$

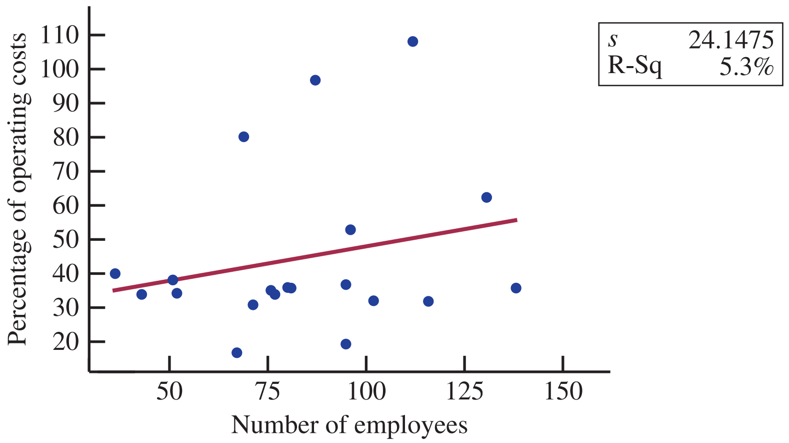
a. For what value of r is s_e as large as s_y? What is the least-squares line in this case?

b. For what values of r will s_e be much smaller than s_y?

Figure for Exercise 5.50

Fitted Line Plot

Percentage of operating costs = 27.71 + 0.2011 number of employees



Section 5.4 Nonlinear Relationships and Transformations

When the points in a scatterplot exhibit a linear pattern and the residual plot does not reveal any problems with the linear fit, the least-squares line is a sensible way to summarize the relationship between two numerical variables. But what should you do if a scatterplot or residual plot exhibits a curved pattern, indicating a more complicated relationship between x and y? When this happens, a nonlinear function can be used to summarize the relationship between x and y. The process is similar to what is done for linear relationships, but now a curve is used instead of a line. After choosing an appropriate curve, a residual plot can be used to assess whether the curve is an appropriate way to describe the relationship. If you decide the curve is a useful summary of the relationship between x and y, you can then use the curve to make predictions.

Choosing a Nonlinear Function to Describe a Relationship

Once you have decided to use a nonlinear function to describe the relationship between two variables x and y, you need to decide what type of function should be considered. Some common functions that are used to model relationships are shown in Figure 5.27.

Example 5.15 Fishing Bears

Understand the context ➤

● The article **"Quantifying Spatiotemporal Overlap of Alaskan Brown Bears and People" (*Journal of Wildlife Management* [2005]: 810–817)** describes a study of the effect of human activity on bears. The researchers wondered if sport fishing and boating might be limiting bears' access to salmon. They collected data on

x = Number of days from the beginning of June (June 1st = 1)

and

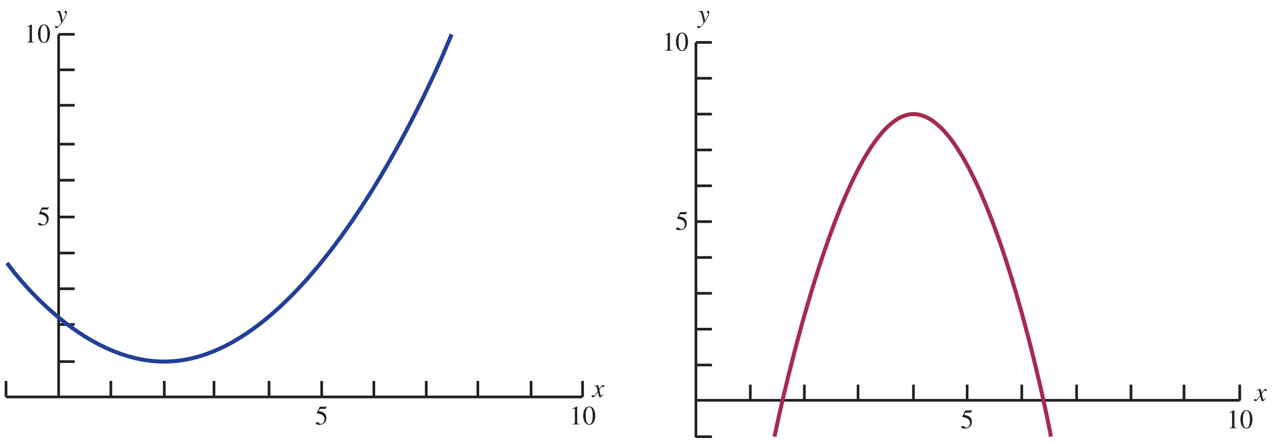
y = Total fishing time (in bear-hours) for the day at a particular location in Alaska

For example, the data pair (33, 45.7) corresponds to the day that is 32 days after June 1 (which is July 2). On this day, the total fishing time for bears was 45.7 bear-hours. This is the sum of the number of hours spent fishing for all bears fishing at this location on July 2.

● **Data set available online**

Figure 5.27
Common nonlinear functions.

	Function	Equation
(a)	Quadratic	$\hat{y} = a + b_1x + x^2$

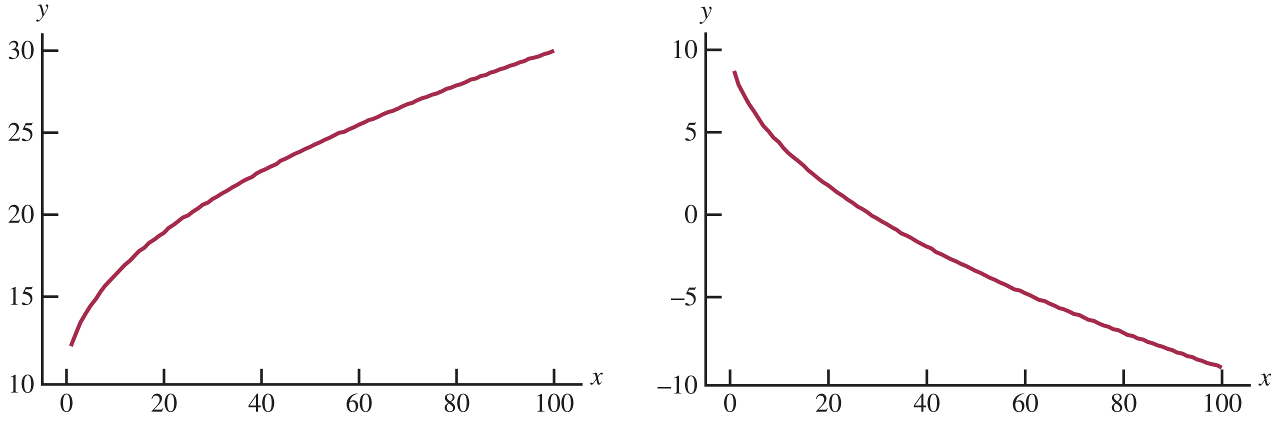


(continued)

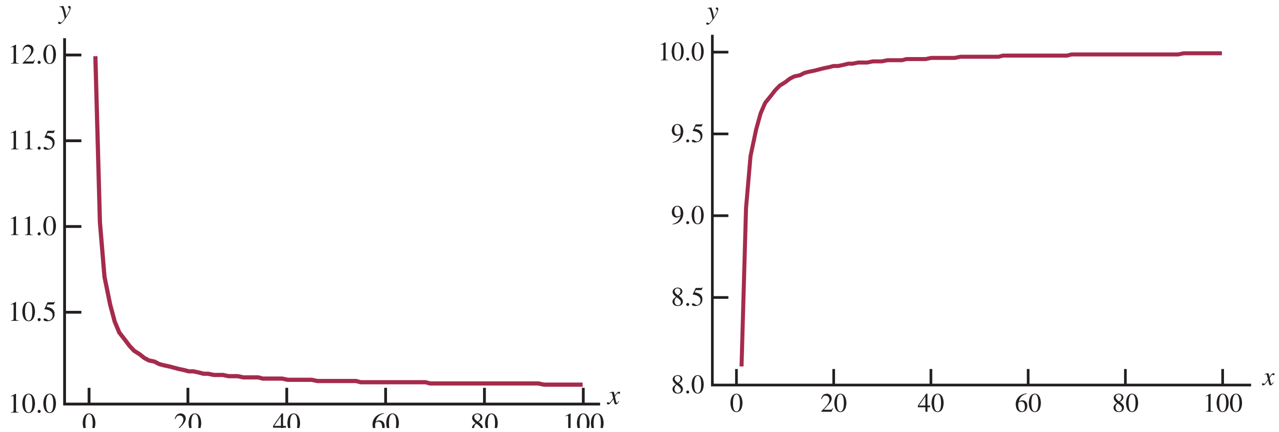
Copyright 2025 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s). Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

Figure 5.27 (continued)
Common nonlinear functions.

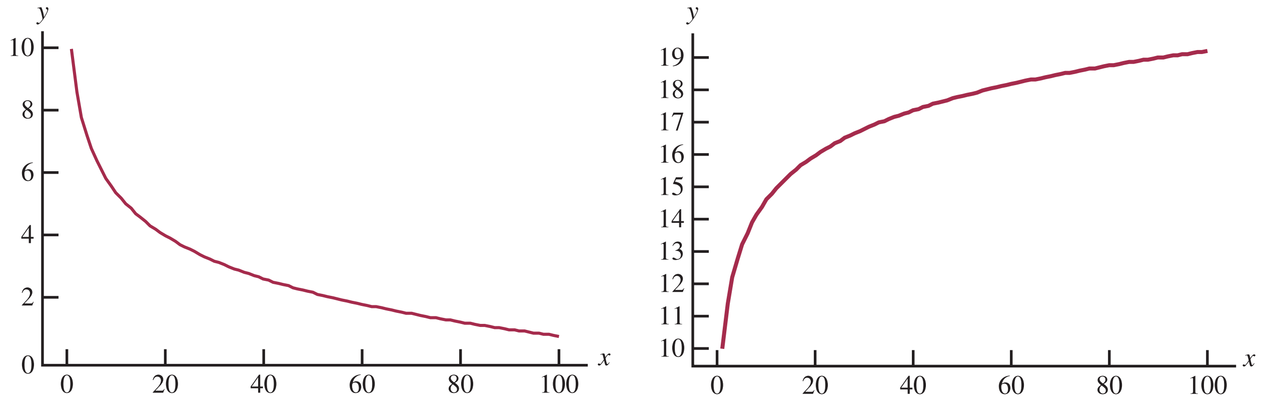
	Function	**Equation**
(b)	Square Root	$\hat{y} = a + b\sqrt{x}$



(c)	Reciprocal	$\hat{y} = a + b\left(\frac{1}{x}\right)$



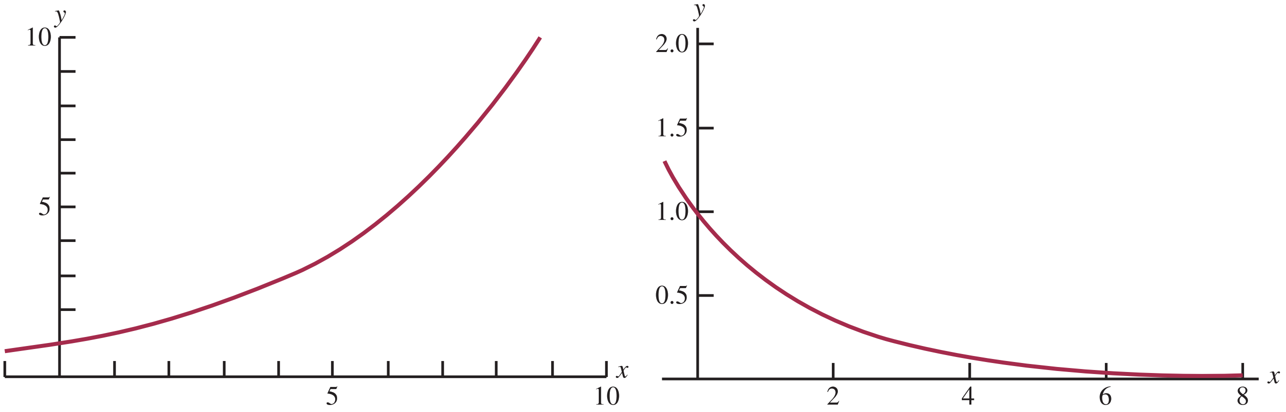
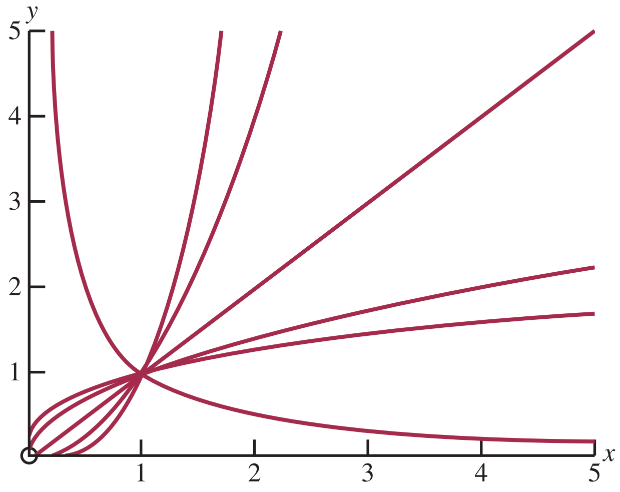
(d)	Log	$\hat{y} = a + b\ln(x)$



(continued)

Copyright 2025 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s). Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

Figure 5.27 (continued)
Common nonlinear functions.

	Function	Equation
(e)	Exponential	$\hat{y} = e^{a+bx}$
		
(f)	Power	$\hat{y} = ax^b$
		

Example 5.15 continued

Consider the data ➤

Date (x) (June 1 = 1)	Fishing Time (bear-hours)	Date (x) (June 1 = 1)	Fishing Time (bear-hours)	Date (x) (June 1 = 1)	Fishing Time (bear-hours)
11	11.3	24	11.3	40	22.0
12	15.1	25	18.4	41	26.0
13	6.6	27	16.2	44	10.5
14	12.9	28	19.5	45	18.6
15	12.1	31	35.8	46	21.1
17	18.1	32	37.1	49	11.9
18	20.9	33	45.7	50	13.7
19	17.6	36	34.8	51	13.7
20	11.0	37	25.6	54	6.3
21	24.6	38	26.7	55	1.8

Figure 5.28 is a scatterplot of these data.

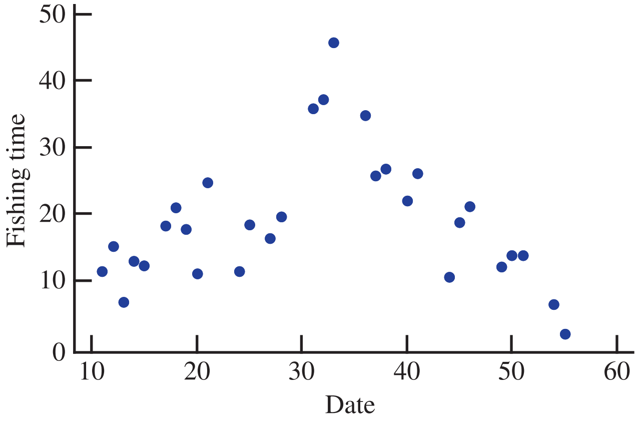


Figure 5.28
Scatterplot of fishing time versus date.

Formulate a plan ➤

The pattern in the scatterplot is clearly nonlinear. Because the shape of the pattern looks like a quadratic curve (see Figure 5.27(a)), a quadratic function could be used to describe this relationship. This suggests considering the following model:

$$\hat{y} = a + b_1x + b_2x^2$$

Once you have decided on a potential nonlinear model, such as quadratic or exponential, the next step is to estimate the coefficients in the model. For example, for the quadratic model $\hat{y} = a + b_1x + b_2x^2$, you would need to use the sample data to determine appropriate values for a, b_1, and b_2. The process used to estimate the coefficients in a nonlinear model is a bit different for quadratic models than for the other nonlinear models in Figure 5.27. Let's begin by considering quadratic models.

Quadratic Regression Models

The general form of the quadratic regression model is $\hat{y} = a + b_1x + b_2x^2$. What are the best choices for the values of a, b_1, and b_2? In fitting a line to data, the principle of least-squares was used to determine the values of the slope and intercept for the "best fit" line. Least-squares can also be used to fit a quadratic function. The deviations, $y - \hat{y}$, are still represented by vertical distances in the scatterplot, but now they are vertical distances from the points to a parabola (the graph of a quadratic function) rather than to a line, as shown in Figure 5.29. Values for the coefficients in the quadratic function are then chosen to make the sum of the squared deviations as small as possible.

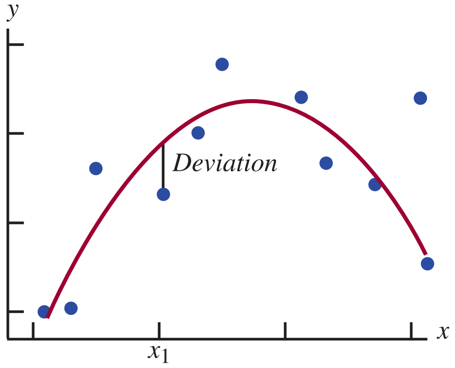


Figure 5.29
Deviation for a quadratic function.

For a quadratic regression model, the least-squares estimates of a, b_1, and b_2 are those values that minimize the sum of squared deviations, $\Sigma(y - \hat{y})^2$, where $\hat{y} = a + b_1x + b_2x^2$.

For quadratic regression, a measure that is useful for assessing fit is

$$R^2 = 1 - \frac{SSResid}{SSTo}$$

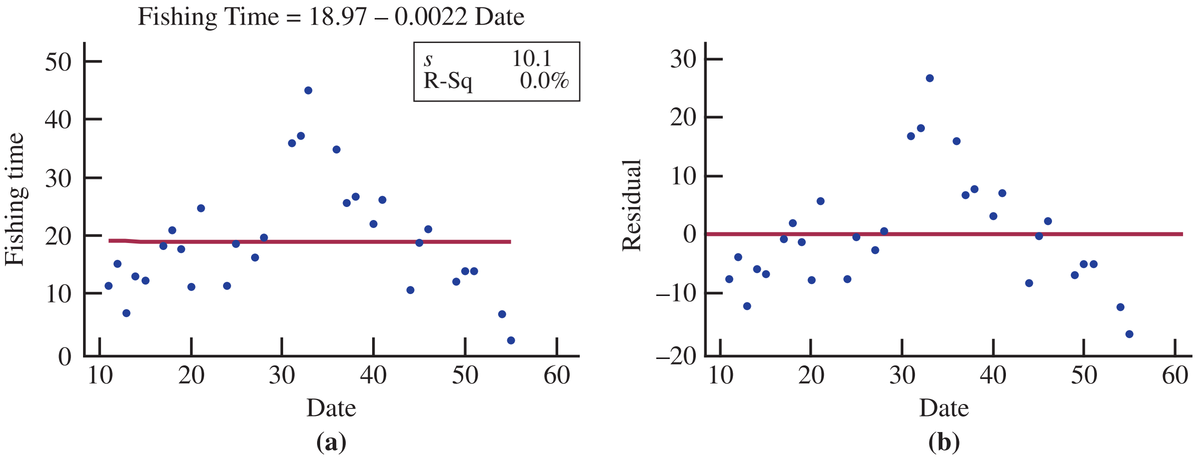
where $SSResid = \Sigma(y - \hat{y})^2$. The measure R^2 is defined in a way similar to r^2 for the linear regression model and is also interpreted in a similar way. The lowercase notation r^2 is used only with linear regression to emphasize the relationship between r^2 and the correlation coefficient, r, in the linear case. For nonlinear models, an uppercase R^2 is used.

The general expressions for calculating the least-squares quadratic regression estimates are somewhat complicated, so you will want to use a statistical software package or a graphing calculator to do the calculations.

Example 5.16 Fishing Bears Revisited—Fitting a Quadratic Model

For the bear fishing data of Example 5.15, the scatterplot showed a curved pattern. If a least-squares line is fit to these data, it is not surprising that the line does not do a good job of describing the relationship ($r^2 = 0.00$, and $s_e = 10.1$). Both the scatterplot and the residual plot show a distinct curved pattern (see Figure 5.30).

Figure 5.30
Plots for the bear fishing data of Examples 5.15 and 5.16:
(a) least-squares line;
(b) residual plot for linear model.



Do the work ➤ JMP output resulting from fitting the least-squares quadratic regression to these data is shown in Figure 5.31.

Figure 5.31
JMP output for the quadratic regression of Example 5.16.

```
Polynomial Fit Degree=2

  Usage = –20.96713 + 2.9957567* Date – 0.0463151* Date^2

Summary of Fit

  RSquare                        0.555541
  RSquare Adj                    0.522618
  Root Mean Square Error         6.856681
  Mean of Response              18.89667
  Observations (or Sum Wgts)    30

Analysis of Variance

Parameter Estimates

  Term        Estimate    Std Error  t Ratio   Prob>[t]
  Intercept  –20.96713    7.567052   –2.77     0.0100*
  Date         2.9957567  0.524223    5.71     <.0001*
  Date^2      –0.046315   0.007973   –5.81     <.0001*
```

From the JMP output, the least-squares quadratic is

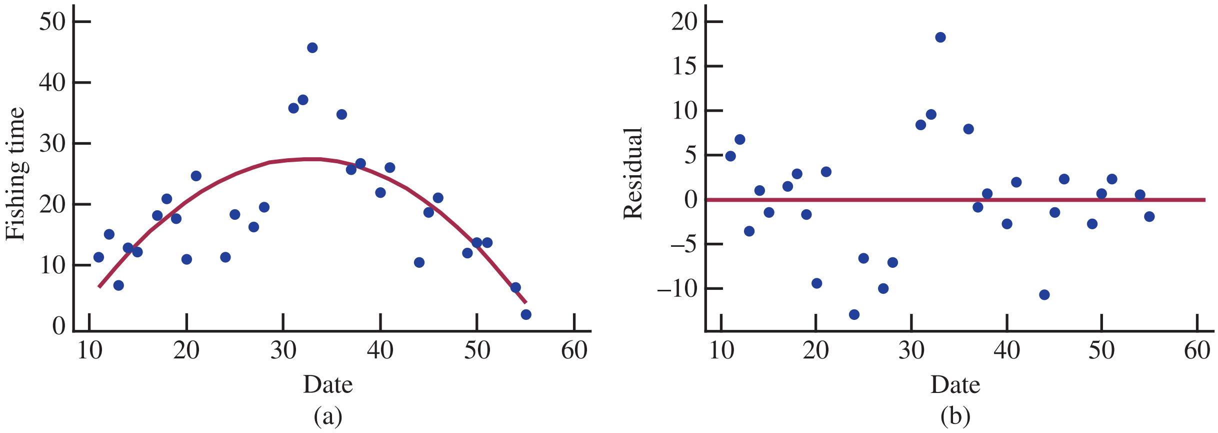
$$\hat{y} = -20.967 + 2.996x - 0.046x^2$$

Copyright 2025 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s). Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

Interpret the results ➤

The curve and the corresponding residual plot for the quadratic regression are shown in Figure 5.32. Notice that there is no strong pattern in this residual plot, unlike the linear case. For the quadratic regression, $R^2 = 0.556$ (as opposed to essentially zero for the linear model). This means that 55.6% of the variability in the bear fishing time can be explained by an approximate quadratic relationship between fishing time and date.

Figure 5.32
Quadratic regression of Example 5.16:
(a) scatterplot with curve;
(b) residual plot for quadratic model.



To use the quadratic regression model to make a prediction, substitute a value of x into the regression equation. For example, the predicted fishing time for June 18 ($x = 18$) is

$$\begin{aligned}\hat{y} &= -20.967 + 2.996x - 0.046x^2 \\ &= -20.967 + 2.996(18) - 0.046(18)^2 \\ &= -20.967 + 53.928 - 14.904 \\ &= 18.057 \text{ bear-hours}\end{aligned}$$

Other Nonlinear Regression Models: Using Transformations

Now that you have seen how to use a quadratic regression model to describe relationships that look like the curves in Figure 5.27(a), consider some other nonlinear models. The square root function of Figure 5.27(b) is

$$\hat{y} = a + b\sqrt{x}$$

Notice that if you define a new variable x' where

$$x' = \sqrt{x}$$

the square root model $\hat{y} = a + b\sqrt{x}$ can be written as $\hat{y} = a + bx'$, so y is a linear function of x'. This suggests that if a scatterplot of (x, y) pairs looks like the curve in Figure 5.26(b), a scatterplot of (x', y) pairs should look linear. You can then use a line to describe the relationship between y and x'—something you already know how to do. This is illustrated in the following example.

Example 5.17 River Water Velocity and Distance from Shore

Understand the context ➤

● Data set available online

● As fans of white-water rafting know, a river flows more slowly close to its banks (because of friction between the riverbank and the water). To study the nature of the relationship between water velocity and the distance from the shore, data were gathered on

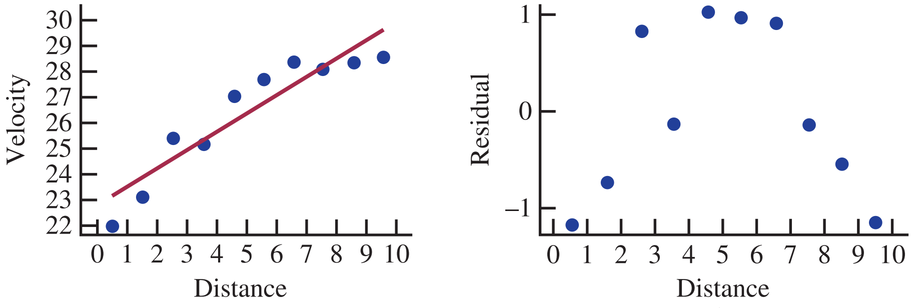
Copyright 2025 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s). Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

velocity (in centimeters per second) of a river at different distances (in meters) from the bank. Suppose that the resulting data were as follows:

Consider the data ➤

Distance	0.5	1.5	2.5	3.5	4.5	5.5	6.5	7.5	8.5	9.5
Velocity	22.00	23.18	25.48	25.25	27.15	27.83	28.49	28.18	28.50	28.63

There is a nonlinear relationship between these two variables, as seen in both the scatterplot (Figure 5.33(a)) and the residual plot from a linear fit (Figure 5.33(b)).



(a) (b)

Figure 5.33 Plots for the data of Example 5.17: (a) scatterplot of the river data; (b) residual plot for linear model.

Formulate a plan ➤

Because the scatterplot of the (x, y) pairs looks like the curve for the square root function in Figure 5.27(b), begin by replacing each x value by its square root:

$$x' = \sqrt{x}$$

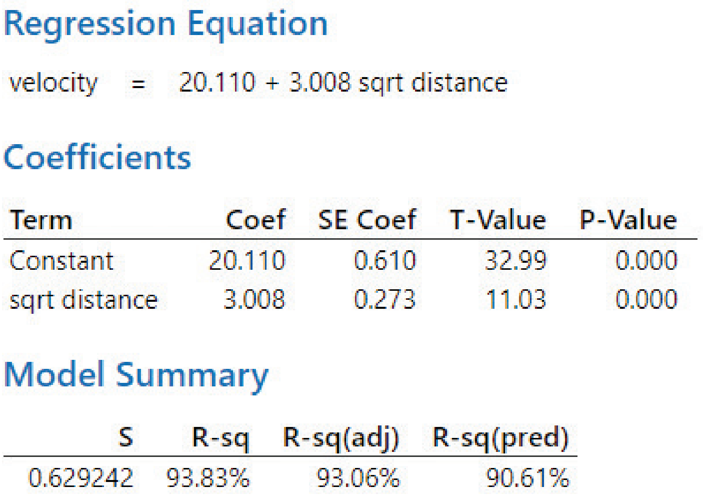
This is called transforming the x values. The original and transformed data are shown in Table 5.2.

Table 5.2 Original and Transformed Data of Example 5.17

Do the work ➤

Original Data		Transformed Data		Original Data		Transformed Data	
x	y	x'	y	x	y	x'	y
0.5	22.00	0.7071	22.00	5.5	27.83	2.3452	27.83
1.5	23.18	1.2247	23.18	6.5	28.49	2.5495	28.49
2.5	25.48	1.5811	25.48	7.5	28.18	2.7386	28.18
3.5	25.25	1.8708	25.25	8.5	28.50	2.9155	28.50
4.5	27.15	2.1213	27.15	9.5	28.63	3.0822	28.63

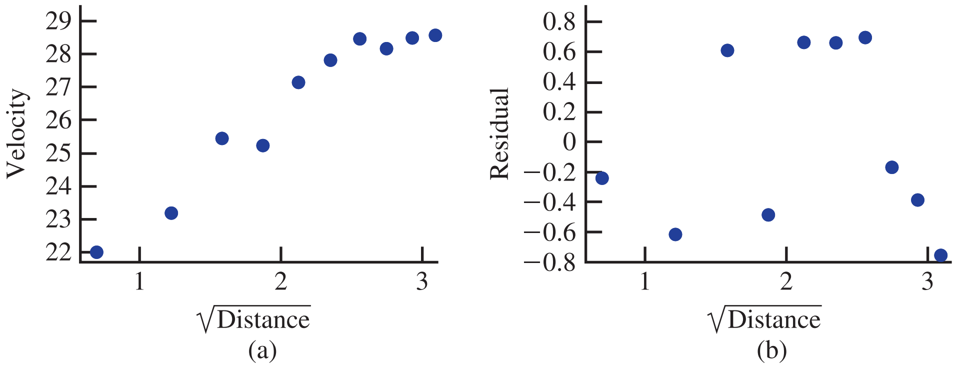
Figure 5.34(a) shows a scatterplot of y versus x' (or equivalently y versus $\sqrt{x}$). The pattern of points in this plot looks linear, and so you can find the least-squares line using the transformed data. The Minitab output using the transformed data is given here:



```
Regression Equation

velocity  =  20.110 + 3.008 sqrt distance

Coefficients

Term             Coef   SE Coef   T-Value   P-Value
Constant       20.110     0.610     32.99     0.000
sqrt distance   3.008     0.273     11.03     0.000

Model Summary

       S     R-sq   R-sq(adj)   R-sq(pred)
0.629242   93.83%      93.06%       90.61%
```

Copyright 2025 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s). Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

Figure 5.34
Plots for the transformed data of Example 5.17:
(a) scatterplot of *y* versus x';
(b) residual plot resulting from a linear fit to the transformed data.



The resulting regression equation is

$$\hat{y} = 20.110 + 3.008x'$$

or, equivalently,

$$\hat{y} = 20.110 + 3.008\sqrt{x}$$

Interpret the results ➤

This model is evaluated using R^2, s_e, and a residual plot. The residual plot (shown in Figure 5.34(b)) does not have any patterns or unusual features. The values of R^2 and s_e (see the Minitab output) indicate that a line is a reasonable way to describe the relationship between *y* and x'.

To predict velocity of the river at a distance of 9 meters from shore, first calculate $x' = \sqrt{x} = \sqrt{9} = 3$ and then use the least-squares line to obtain a predicted *y* value:

$$\hat{y} = 20.110 + 3.008x' = 20.110 + (3.008)(3) = 29.134 \text{ cm per second}$$

One strategy for fitting a nonlinear model is to find a way to transform the *x* and/or *y* values so that a scatterplot of the transformed data shows a linear pattern. A **transformation** (sometimes called a re-expression) involves using a simple function of a variable in place of the variable itself (like $x' = \sqrt{x}$ in Example 5.17). Common transformations involve taking square roots, logarithms, or reciprocals.

The nonlinear models (other than the quadratic model) given in Figure 5.27 (square root, reciprocal, log, exponential, and power) can all be fit by transforming variables and then fitting a line to the transformed data. It is convenient to separate these models into two types:

1. Models that involve transforming *only* the *x* variable (square root, reciprocal, and log).
2. Models that involve transforming the *y* variable (exponential and power).

Models That Involve Transforming Only *x*

The square root, reciprocal, and log models all have the form

$$\hat{y} = a + b(\text{some function of } x)$$

where the function of *x* is square root, reciprocal, or log. Notice that these models all describe a linear relationship between *y* and a function of *x*. This suggests that if the pattern in the scatterplot of (*x*, *y*) pairs looks like one of the curves in Figure 5.27(b)–(d), an appropriate transformation of the *x* values should result in transformed data that shows a linear pattern.

Model	**Transformation**
Square root	$x' = \sqrt{x}$
Reciprocal	$x' = \frac{1}{x}$
Log	$x' = \ln(x)$

Copyright 2025 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s). Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

Example 5.18 Electromagnetic Radiation

Understand the context ➤

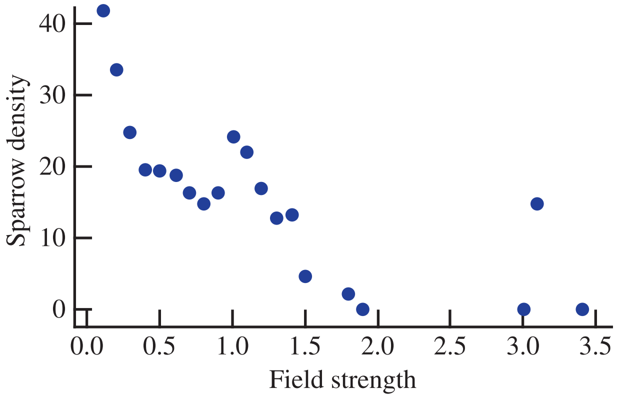
● Is electromagnetic radiation from phone antennae associated with declining bird populations? This is one of the questions investigated by the authors of the paper **"The Urban Decline of the House Sparrow (*Passer domesticus*): A Possible Link with Electromagnetic Radiation" (*Electromagnetic Biology and Medicine* [2007]: 141–151)**. The accompanying data on x = Field strength (the strength of the electromagnetic field in volts per meter) and y = Sparrow density (sparrows per hectare) were read from a graph in the paper.

Consider the data ➤

Field Strength	Sparrow Density	Field Strength	Sparrow Density
0.11	41.71	0.90	16.29
0.20	33.60	1.20	16.97
0.29	24.74	1.30	12.83
0.40	19.50	1.41	13.17
0.50	19.42	1.50	4.64
0.61	18.74	1.80	2.11
1.01	24.23	1.90	0.00
1.10	22.04	3.01	0.00
0.70	16.29	3.10	14.69
0.80	14.69	3.41	0.00

A scatterplot of these data is shown in Figure 5.35.

Figure 5.35
Scatterplot of the sparrow data.



Formulate a plan ➤

Notice that there is a curved pattern in the scatterplot that has a shape that is similar to the first graph in Figure 5.27(d). This suggests that a reasonable choice for describing the relationship between sparrow density and field strength is a log model

$$\hat{y} = a + b\ln(x)$$

To fit this model, begin by transforming the x values using

$$x' = \ln(x)$$

Do the work ➤

In this example, natural logs (denoted by ln or $\log_e$) are used, but in practice, either the natural log or log base 10 (denoted by log or $\log_{10}$) can be used. The transformed data are given here.

ln(Field Strength)	Sparrow Density	ln(Field Strength)	Sparrow Density
−2.207	41.71	−0.105	16.29
−1.609	33.60	0.182	16.97
−1.238	24.74	0.262	12.83
−0.916	19.50	0.344	13.17
−0.693	19.42	0.405	4.64
−0.494	18.74	0.588	2.11
0.001	24.23	0.642	0.00
0.095	22.04	1.102	0.00
−0.357	16.29	1.131	14.69
−0.223	14.69	1.227	0.00

● **Data set available online**

Copyright 2025 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s). Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

A scatterplot of the (x', y) pairs is shown in Figure 5.36. The pattern in this plot looks linear, so you can find the least-squares line using the transformed data. From the resulting Minitab output (Figure 5.37), the equation of the least-squares line is

$$\hat{y} = 14.8 - 10.5x'$$

or, in terms of x

$$\hat{y} = 14.8 - 10.5\ln(x)$$

Figure 5.36
Scatterplot of the transformed data of Example 5.18.

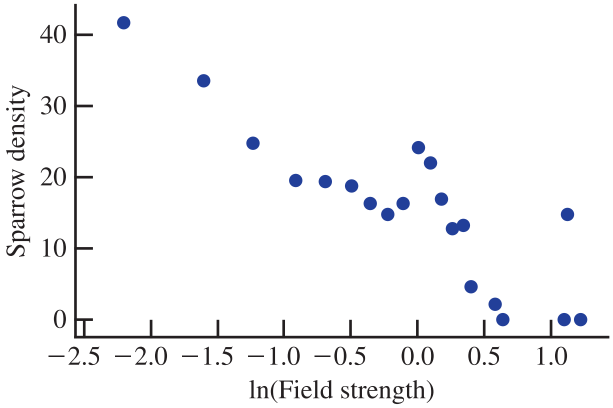


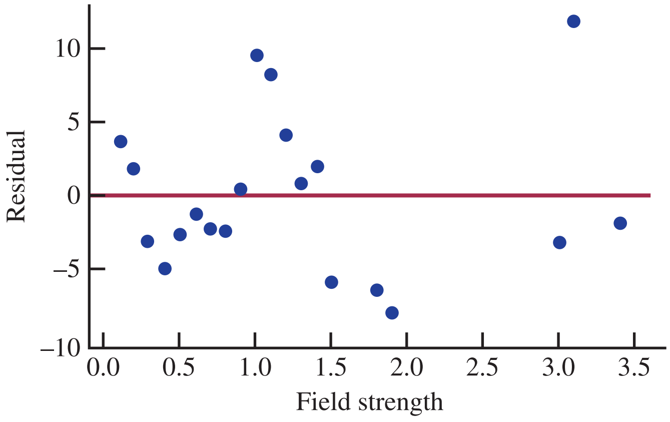
Figure 5.37
Minitab output for the log model of Example 5.18.

```
Regression Analysis: Sparrow Density Versus ln(Field Strength)

The regression equation is
Sparrow Density = 14.8 – 10.5 ln(Field Strength)

Predictor              Coef    SE Coef        T        P
Constant             14.805      1.238    11.96    0.000
ln (Field Strength) –10.546      1.389    –7.59    0.000

S = 5.50641            R–Sq = 76.2%          R–Sq(adj) = 74.9%
```

Interpret the results ➤

The value of R^2 for this model is 0.762 and $s_e = 5.5$. A residual plot from the least-squares line fit to the transformed data is shown in Figure 5.38. There are no apparent patterns or unusual features in the residual plot, so it appears that the log model is a reasonable choice for describing the relationship between sparrow density and field strength.

Figure 5.38
Residual plot for the log model fit to the sparrow data of Example 5.18.



Copyright 2025 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s). Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

This model can now be used to predict sparrow density from field strength. For example, if the field strength is 1.6 Volts per meter, you would predict that sparrow density would be

$$\begin{aligned}\hat{y} &= 14.8 - 10.5x' \\ &= 14.8 - 10.5 \ln(x) \\ &= 14.8 - 10.5 \ln(1.6) \\ &= 14.8 - 10.5\,(0.470) \\ &= 9.685 \text{ sparrows per hectare}\end{aligned}$$

Models That Involve Transforming *y*

There are two nonlinear models in Figure 5.27 that have not yet been considered—the exponential and power models. Let's begin by considering the exponential model

$$y = e^{a+bx}$$

Using properties of logarithms, it follows that

$$\begin{aligned}\ln(y) &= \ln e^{a+bx} \\ &= a + bx\end{aligned}$$

For the exponential model, ln(y) is a linear function of x. This means that if the pattern in a scatterplot of the (x, y) pairs looks like the exponential curves of Figure 5.27(e), a scatterplot of the (x, y') pairs, where

$$y' = \ln y$$

should show a linear pattern.

Now consider the power model of Figure 5.27(f),

$$y = ax^b$$

Again using properties of logarithms,

$$\begin{aligned}\ln y &= \ln(ax^b) \\ &= \ln a + \ln x^b \\ &= \ln a + b\ln x\end{aligned}$$

If a power model is an appropriate way to describe the relationship between x and y, a linear model would describe the relationship between $x' = \ln x$ and $y' = \ln y$.

Model	**Transformation**
Exponential	$y' = \ln y$
Power	$x' = \ln x$ $y' = \ln y$

Example 5.19 Loons in Acidic Lakes

Understand the context ➤

● A study of factors that affect the survival of loon chicks is described in the paper **"Does Prey Biomass or Mercury Exposure Affect Loon Chick Survival in Wisconsin?" (*The Journal of Wildlife Management* [2005]: 57–67)**. In this study, a relationship between the pH of lake water and blood mercury level in loon chicks was observed.

Consider the data ➤

The researchers thought that it is possible that the pH of the lake water could be affecting the type of fish that the loons ate. The accompanying data (Table 5.3) on x = Lake pH and y = Blood mercury level (mg/g) for 37 loon chicks from different lakes in Wisconsin were read from a graph in the paper. A scatterplot is shown in Figure 5.39(a).

● **Data set available online**

Copyright 2025 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s). Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

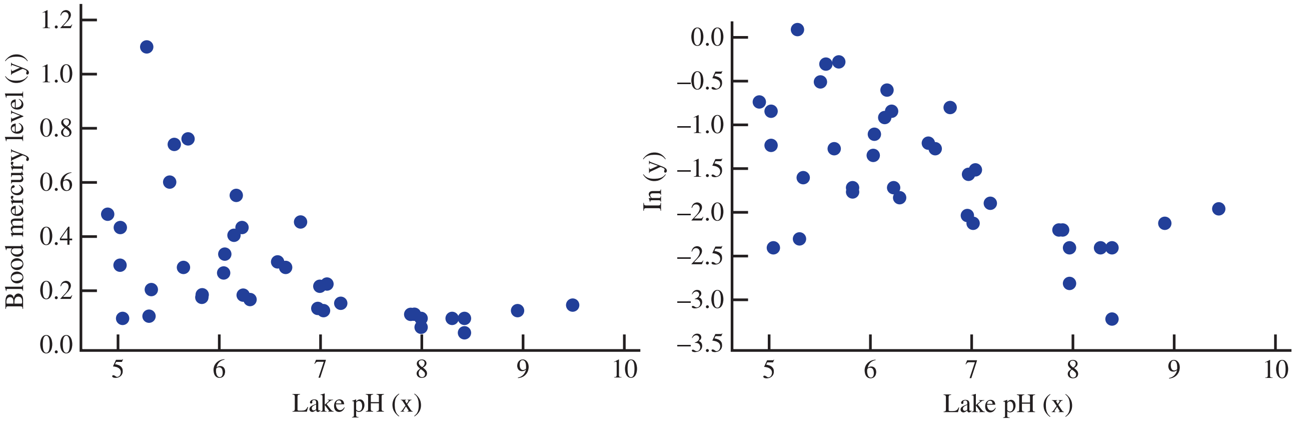
Table 5.3 Data and Transformed Data from Example 5.19

Lake pH (x)	Blood Mercury Level (y)	$y' = \ln(y)$	Lake pH (x)	Blood Mercury Level (y)	$y' = \ln(y)$	Lake pH (x)	Blood Mercury Level (y)	$y' = \ln(y)$
5.28	1.10	0.0953	6.17	0.55	−0.5978	7.03	0.12	−2.1203
5.69	0.76	−0.2744	6.22	0.43	−0.8440	7.20	0.15	−1.8971
5.56	0.74	−0.3011	6.15	0.40	−0.9163	7.89	0.11	−2.2073
5.51	0.60	−0.5108	6.05	0.33	−1.1087	7.93	0.11	−2.2073
4.90	0.48	−0.7340	6.04	0.26	−1.3471	7.99	0.09	−2.4079
5.02	0.43	−0.8440	6.24	0.18	−1.7148	7.99	0.06	−2.8134
5.02	0.29	−1.2379	6.30	0.16	−1.8326	8.30	0.09	−2.4079
5.04	0.09	−2.4079	6.80	0.45	−0.7985	8.42	0.09	−2.4079
5.30	0.10	−2.3026	6.58	0.30	−1.2040	8.42	0.04	−3.2189
5.33	0.20	−1.6094	6.65	0.28	−1.2730	8.95	0.12	−2.1203
5.64	0.28	−1.2730	7.06	0.22	−1.5141	9.49	0.14	−1.9661
5.83	0.17	−1.7720	6.99	0.21	−1.5606			
5.83	0.18	−1.7148	6.97	0.13	−2.0402			

Formulate a plan ➤

The pattern in the scatterplot is typical of exponential decay and looks like the second curve in Figure 5.27(e). The change in y as x increases is much smaller for large x values than for small x values. You can see that a change of 1 in pH is associated with a much larger change in blood mercury level in the part of the plot where the x values are small than in the part of the plot where the x values are large.

Figure 5.39
Scatterplots for the loon data of Example 5.19:
(a) scatterplot of original data;
(b) scatterplot of transformed data with $y' = \ln y$.



(a)

(b)

To fit an exponential model, begin by transforming the y values using $y' = \ln y$. The transformed y values are given in Table 5.3. A scatterplot of the (x, y') pairs is shown in Figure 5.39(b). The pattern in the scatterplot of the transformed data looks linear, so it is reasonable to use the least-squares line to describe the relationship between y' and x.

The following Minitab output is the result of fitting the least-squares line to the transformed data:

Do the work ➤

```
The regression equation is

ln(y) = 1.06 − 0.396 Lake pH (x)

Predictor            Coef     SE Coef       T        P
Constant           1.0550      0.5535    1.91    0.065
Lake pH (x)      −0.39564     0.08264   −4.79    0.000

S = 0.605645       R-Sq = 39.6%        R-Sq(adj) = 37.8%
```

Copyright 2025 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s). Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

The resulting least-squares line is

$$\hat{y}' = 1.06 - 0.396x$$

or equivalently

$$\widehat{\ln y} = 1.06 - 0.396x$$

Notice that the linear model in Example 5.19 could be used to predict values of ln(y). To translate this linear model back to the form of the exponential model, you can take antilogarithms of each side of the equation. This is illustrated in Example 5.20.

Example 5.20 Revisiting the Loon Data

For the loon data of Example 5.19, $y' = \ln y$ and the equation of the least-squares line describing the relationship between y' and x was $\hat{y}' = 1.06 - 0.396x$ or $\widehat{\ln y} = 1.06 - 0.396x$. To "undo" the transformation, take the antilog of both sides of this equation:

$$e^{\ln y} = e^{1.06 - 0.396x}$$

Using properties of logs and exponents, $e^{\ln y} = e^{1.06 - 0.396x} = (e^{1.06})\,(e^{-0.396x})$. Then

$$\hat{y} = e^{1.06}\,e^{-0.396x} = 2.886e^{-0.396x}$$

The equation $\hat{y} = 2.886e^{-0.396x}$ can be used to predict the y value (blood mercury level) for a given x (lake pH). For example, the predicted blood mercury level when lake pH is 6 is

$$\hat{y} = 2.886e^{-0.396x} = 2.886e^{0.396(6)} = 2.886e^{-2.376} = 2.886(0.093) = 0.268$$

The process of transforming data, fitting a line to the transformed data, and then undoing the transformation to get an equation for a curved relationship between x and y usually results in a curve that provides a reasonable fit to the sample data. But when y is transformed, this process does not result in the least-squares curve for the data. For example, in Example 5.20, a transformation was used to fit the curve $\hat{y} = 2.886e^{-0.396x}$. However, there may be another equation of the form $\hat{y} = ae^{bx}$ that has a smaller sum of squared residuals for the *original* data than the one obtained using transformations. Finding the least-squares estimates for a and b for an exponential or power model is difficult. Fortunately, the curves found using transformations usually provide reasonable predictions of y.

Choosing Among Different Possible Nonlinear Models

Often there is more than one reasonable model that could be used to describe a nonlinear relationship between two variables. When this is the case, how do you choose a model? The choice often involves considering how well the model fits the data and what scientific theory suggests the form of the relationship might be. In the absence of any scientific theory, you want to choose a model that has small residuals (small s_e) and accounts for a large proportion of the variability in y (large R^2). On the other hand, if scientific theory suggests that the relationship should have a particular form, you would choose a model that is consistent with the suggested form. The following two examples illustrate these two different situations.

Example 5.21 Look for Salamanders Under Those Rocks

Understand the context ➤

The paper **"The Relationship Between Rock Density and Salamander Density in a Mountain Stream" (*Herpetologica* [1987]: 357–361)** describes a study of factors that influence population density of salamanders. In this study, researchers created a range of salamander habitats at different locations in a small stream in the Appalachian Mountains by using different sized rocks and pebbles. Three months later, they returned to measure salamander density. This resulted in data on

x = Rock density (rocks per 1.4 square meters)

and

$$y = \text{Salamander density (salamanders per 1.4 square meters)}$$

Consider the data ➤ A scatterplot of these data is shown in Figure 5.40.

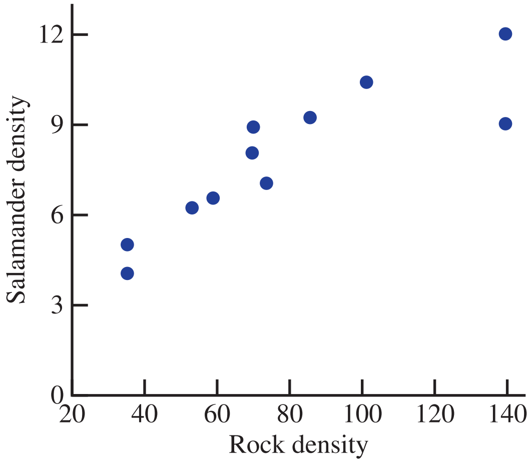


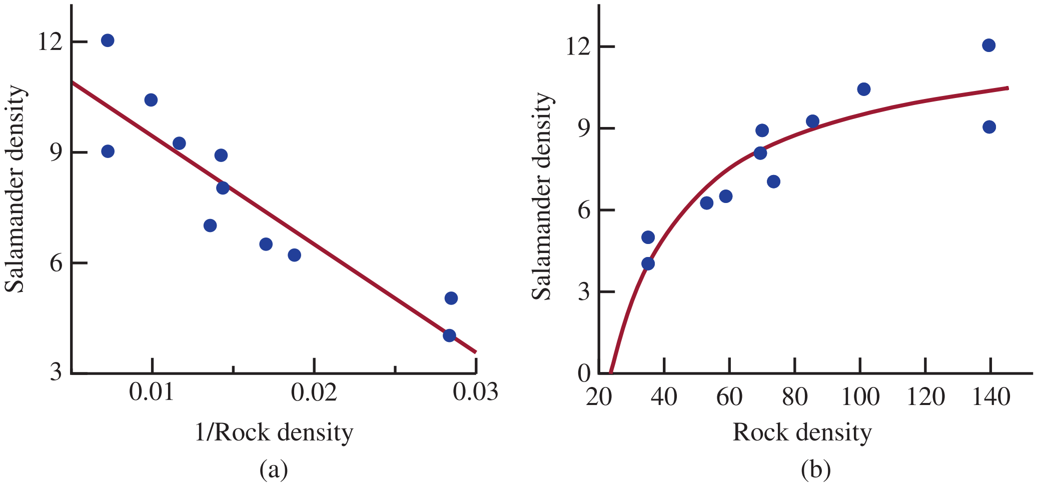
Figure 5.40
Scatterplot of salamander density versus rock density.

Formulate a plan ➤ Looking at the curves in Figure 5.27, you might think about using a square root model or a reciprocal model. The researchers chose the reciprocal regression model, $y = a + b\left(\frac{1}{x}\right)$, for two reasons: (1) it provided a good fit to the data and (2) the reciprocal model made sense scientifically. The investigators felt that there would be an upper limit to the population density, since the streambed is a nonrenewable resource and the stream therefore had a limit on the number of salamanders that could be sustained. This limit is known as the "carrying capacity" of an environment and is estimated by the value of the intercept, *a*, in the reciprocal model.

Do the work ➤ Using the transformation $x' = \frac{1}{x}$, a least-squares line was fit using the transformed data. The resulting model was

$$y = 12.37 - 292.6x' = 12.37 - 292.6\left(\frac{1}{x}\right)$$

The value of R^2 for this model is $R^2 = 0.82$. Figure 5.41(a) shows a scatterplot of the transformed data and the least-squares line. Figure 5.41(b) is a scatterplot of the original data that also shows the curve corresponding to the reciprocal model.



(a)

(b)

Figure 5.41
Scatterplots for the salamander data of Example 5.21:
(a) transformed data;
(b) original data.

Copyright 2025 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s). Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

Example 5.22 How Old Is That Lobster?

Understand the context ➤

● Can you tell how old a lobster is by its size? This question was investigated by the authors of a paper that appeared in the ***Biological Bulletin* (August, 2007)**. Researchers measured carapace (the exterior shell) length (in mm) of 27 laboratory-raised lobsters of known age. The data on x = Carapace length and y = Age (in years) in Table 5.4 were read from a graph in the paper.

Table 5.4 Original and Transformed Data for Example 5.22

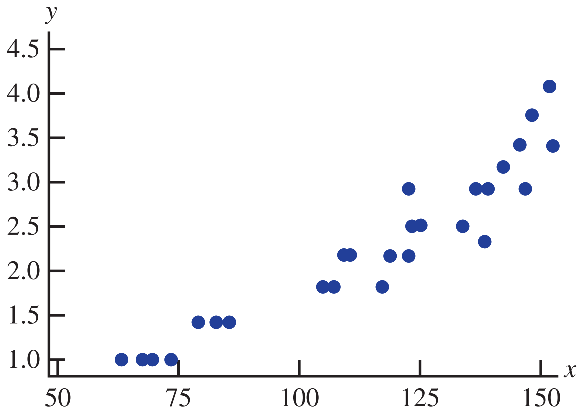
Consider the data ➤

y	x	ln(x)	ln(y)	y	x	ln(x)	ln(y)
1.00	63.32	4.148	0.000	2.33	138.47	4.931	0.846
1.00	67.50	4.212	0.000	2.50	133.95	4.897	0.916
1.00	69.58	4.242	0.000	2.51	125.25	4.830	0.920
1.00	73.41	4.296	0.000	2.50	123.51	4.816	0.916
1.42	79.32	4.373	0.351	2.93	146.82	4.989	1.075
1.42	82.80	4.416	0.351	2.92	139.17	4.936	1.072
1.42	85.59	4.450	0.351	2.92	136.73	4.918	1.072
1.82	105.07	4.655	0.599	2.92	122.81	4.811	1.072
1.82	107.16	4.674	0.599	3.17	142.30	4.958	1.154
1.82	117.25	4.764	0.599	3.41	152.73	5.029	1.227
2.18	109.24	4.694	0.779	3.42	145.78	4.982	1.230
2.18	110.64	4.706	0.779	3.75	148.21	4.999	1.322
2.17	118.99	4.779	0.775	4.08	152.04	5.024	1.406
2.17	122.81	4.811	0.775				

Formulate a plan ➤

The scatterplot of the data in Figure 5.42 shows a clear curved pattern. Considering the curves in Figure 5.27, there are several models that could be considered, including the exponential model and the power model.

Figure 5.42
Scatterplot of age versus carapace length.



In this situation, there is no scientific theory to suggest what the form of the relationship might be. A reasonable way to proceed is to fit both models and then choose the one that provides the best fit. Let's start with the exponential model. Using the transformation

$$y' = \ln y$$

results in the transformed y values given in Table 5.4. A scatterplot of the (x, y') pairs is shown in Figure 5.43(a). Notice that the relationship between y' and x looks linear.

● **Data set available online**

Copyright 2025 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s). Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

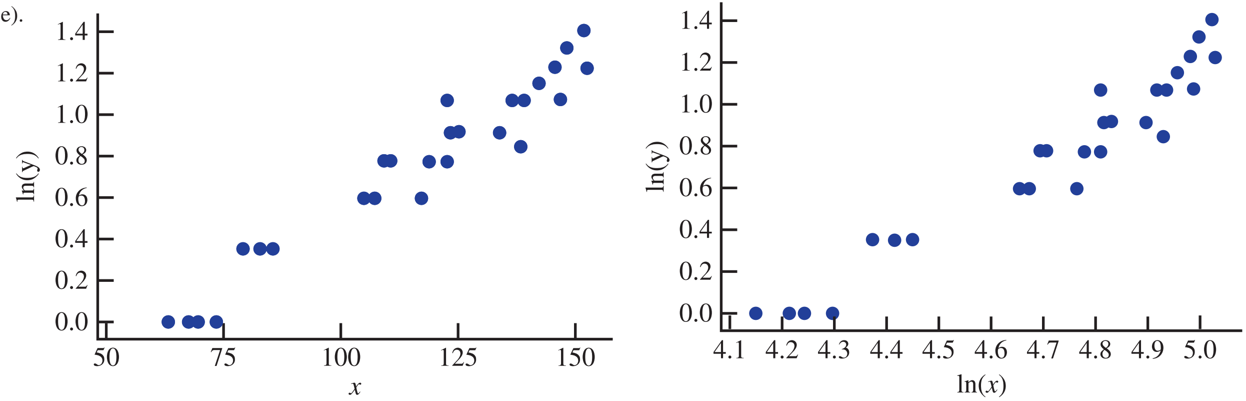
Minitab was used to fit the least-squares line to the (x, y') data, resulting in the following output:

Do the work ➤

Regression Analysis: ln(*y*) versus *x*

```
The regression equation is
ln(y) = −0.927 + 0.0145 x
Predictor        Coef      SE Coef        T        P
Constant    −0.92704      0.08692    −10.67    0.000
x           0.0144903    0.0007311    19.82    0.000

S = 0.105917          R-Sq = 94.0%          R-Sq(adj) = 93.8%
```

Figure 5.43
Scatterplot of the transformed lobster data:
(a) x = carapace length, y' = ln(age);
(b) x' = ln(carapace length), y' = ln(age).



(a)

(b)

Figure 5.43(b) shows a scatterplot of the transformed data that would be used to fit a power model ($x' = \ln(x)$, $y' = \ln(y)$). The pattern in this scatterplot also looks linear. Minitab output from fitting a least-squares line using the (x', y') data is shown here:

Regression Analysis: ln(*y*) versus ln(*x*)

```
The regression equation is
ln(y) = –6.32 + 1.50ln(x)
Predictor        Coef     SE Coef       T       P
Constant     –6.3205      0.3687   –17.14   0.000
ln(x)        1.49866     0.07805    19.20   0.000

S = 0.109120          R-Sq = 93.6%          R-Sq(adj) = 93.4%
```

Notice that both the exponential model and the power model have large R^2 values. Comparing the fit of the two models results in

Model	R^2	S_e
Exponential	0.940	0.106
Power	0.936	0.109

Interpret the results ➤

Both models do a good job in describing the relationship between age and carapace length. In the absence of any scientific theory suggesting the form of the model, you might choose the exponential model because it provides a (slightly) better fit.

The exponential model is

$$\hat{y}' = -0.927 + 0.0145x$$

or equivalently

$$\ln \hat{y} = -0.927 + 0.0145x$$

or

$$\hat{y} = e^{-0.927}e^{0.0145x} = 0.396e^{0.0145x}$$

Copyright 2025 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s). Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

Exercises 5.54–5.63

● **Data set available online**

5.54 ● The following data on x = Frying time (in seconds) and y = Moisture content (%) appeared in the paper **"Thermal and Physical Properties of Tortilla Chips as a Function of Frying Time" (*Journal of Food Processing and Preservation* [1995]: 175 –189)**:

x	5	10	15	20	25	30	45	60
y	16.3	9.7	8.1	4.2	3.4	2.9	1.9	1.3

a. Construct a scatterplot of the data.
b. Would a line provide an effective summary of the relationship? Explain.

5.55 ● Use the information provided in the previous exercise to answer the following questions.
a. Here are the values of $x' = \log(x)$ and $y' = \log(y)$:

x'	0.70	1.00	1.18	1.30	1.40	1.48	1.65	1.78
y'	1.21	0.99	0.91	0.62	0.53	0.46	0.28	0.11

Construct a scatterplot of these transformed data, and comment on the pattern.
b. Based on the accompanying MINITAB output, does the least-squares line effectively summarize the relationship between y' and x'?

```
The regression equation is
log(moisture) = 2.02 − 1.05 log(time)

Predictor      Coef     SE Coef      T      P
Constant     2.01780   0.09584   21.05  0.000
log(time)   −1.05171   0.07091  −14.83  0.000

S = 0.0657067 R-Sq = 97.3% R-Sq(adj) =
96.9%
```

c. Use the MINITAB output to predict moisture content when frying time is 35 sec.

5.56 ● The paper **"Aspects of Food Finding by Wintering Bald Eagles" (*The Auk* [1983]: 477–484)** examined the relationship between the time that eagles spend aloft searching for food (indicated by the percentage of eagles soaring) and relative food availability. The accompanying data were taken from a scatterplot that appeared in this paper. Salmon availability is denoted by x and the percentage of eagles in the air is denoted by y.

x	0	0	0.2	0.5	0.5	1.0	1.2	1.9	2.6	3.3	4.7	6.5
y	28.2	69.0	27.0	38.5	48.4	31.1	26.9	8.2	4.6	7.4	7.0	6.8

a. Draw a scatterplot for this data set. Would you describe the pattern in the plot as linear or curved?
b. One possible transformation that might lead to a linear pattern involves taking the square root of both the x and y values. Construct a scatterplot using the variables $\sqrt{x}$ and $\sqrt{y}$. Is a line a more reasonable chioce for describing the pattern in this scatterplot than the pattern in the scatterplot in Part (a)?
c. After considering the scatterplot in Part (a), suggest another transformation that might be used to straighten the original plot. Use your suggestion to construct a scatterplot for your x' and/or y'. Which transformation (the one you suggested or the square root transformation from Part (b)) is preferable? Explain.

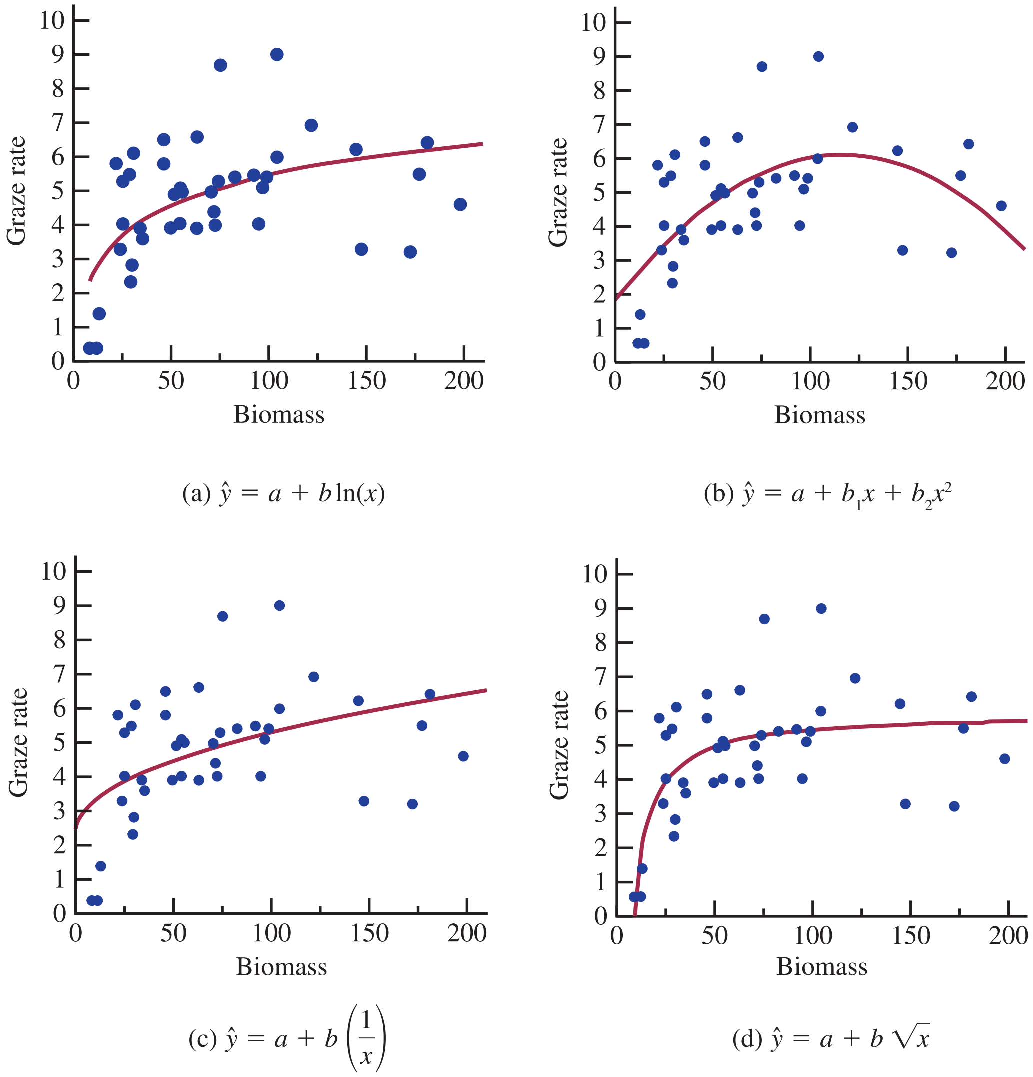
5.57 Food intake of grazing animals is limited by the rate grass can be chewed and swallowed, as well as the rate at which food can be digested. The authors of the paper **"What Constrains Daily Intake in Thomson's Gazelles?" (*Ecology* [1999]: 2338–2347)** observed the grazing activity of captive Thomson's gazelles. They recorded grazing rate (amount of grass eaten, in grams per minute) and biomass of the grazing area (food density, in grams per square meter). Scatterplots of these data with four possible functions that might be used to describe the relationship between grazing rate and biomass are shown in the figure at the top of the following page. Which of these functions would you recommend? Explain your reasoning in a few sentences. (Hint: See Example 5.21.)

5.58 ● A study, described in the paper **"Prediction of Defibrillation Success from a Single Defibrillation Threshold Measurement" (*Circulation* [1988]: 1144–1149)** investigated the relationship between defibrillation success and the energy of the defibrillation shock (expressed as a multiple of the defibrillation threshold). The accompanying data are from this study.

Energy of Shock	Success (%)
0.5	33.3
1.0	58.3
1.5	81.8
2.0	96.7
2.5	100.0

a. Construct a scatterplot of y = Success and x = Energy of shock. Does the relationship appear to be linear or nonlinear?
b. Fit a least-squares line to the given data, and construct a residual plot. Does the residual plot support your conclusion in Part (a)? Explain.

Figure for Exercise 5.57



(a) $\hat{y} = a + b\ln(x)$

(b) $\hat{y} = a + b_1x + b_2x^2$

(c) $\hat{y} = a + b\left(\frac{1}{x}\right)$

(d) $\hat{y} = a + b\sqrt{x}$

5.59 Refer to the data given in the previous exercise.

a. Consider transforming the data by leaving y unchanged and using either $x' = \sqrt{x}$ or $x'' = \ln(x)$. Which of these transformations would you recommend? Justify your choice by appealing to appropriate graphical displays.

b. Using the transformation you recommended in Part (a), find the equation of the least-squares line that describes the relationship between y and the transformed x.

c. What would you predict success to be when the energy of shock is 1.75 times the threshold level? When it is 0.8 times the threshold level?

5.60 ● The following table gives the number of heart transplants performed in the United States each year from 2006 to 2015 **(U.S. Department of Health and Human Services, optn .transplant.hrsa.gov)**:

Year	Number of Heart Transplants
1 (2006)	2,193
2	2,209
3	2,163
4	2,211
5	2,332
6	2,322
7	2,378
8	2,531
9	2,655
10 (2015)	2,804

a. Construct a scatterplot of these data. Describe how the number of heart transplants has changed over time from 2006 to 2015.

b. Find the equation of the least-squares line that describes the relationship between y = Number of heart transplants performed and x = Year.

Copyright 2025 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s). Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

c. Calculate the 10 residuals and construct a residual plot.

d. Are there any features of the residual plot that indicate that the relationship between year and number of heart transplants performed would be better described by curve rather than a line? Explain.

5.61 Refer to the heart transplant data given in the previous exercise.

a. Find a transformation of x and/or y that straightens the plot. Construct a scatterplot for your transformed variables.

b. Using the transformed variables from Part (a), fit a least-squares line and use it to predict the number waiting for a heart transplant in 2016 (Year 11).

c. The prediction made in Part (b) involves prediction for an x value that is outside the range of the x values in the sample. What assumption must you be willing to make for this to be reasonable? Do you think this assumption is reasonable in this case? Would your answer be the same if the prediction had been for the year 2026 rather than 2016? Explain.

5.62 ● The paper **"Population Pressure and Agricultural Intensity" (*Annals of the Association of American Geographers* [1977]: 384–396)** reported a positive relationship between population density and agricultural intensity. The following data consist of measures of population density (x) and agricultural intensity (y) for 18 different subtropical locations:

x	1.0	26.0	1.1	101.0	14.9	134.7
y	9	7	6	50	5	100

x	3.0	5.7	7.6	25.0	143.0	27.5
y	7	14	14	10	50	14

x	103.0	180.0	49.6	140.6	140.0	233.0
y	50	150	10	67	100	100

a. Construct a scatterplot of y versus x. Is the scatterplot compatible with the statement of a positive relationship made in the paper?

b. Construct a scatterplot that uses y and x^2. Does this transformation result in a linear pattern in the scatterplot?

c. Draw a scatterplot that uses log(y) and x. The log(y) values, given in order corresponding to the y values, are 0.95, 0.85, 0.78, 1.70, 0.70, 2.00, 0.85, 1.15, 1.15, 1.00, 1.70, 1.15, 1.70, 2.18, 1.00, 1.83, 2.00, and 2.00. How does this scatterplot compare with that of Part (b)?

d. Now consider a scatterplot that uses transformations on both x and y: log(y) and x^2. Is this effective in creating a linear pattern in the plot? Explain.

5.63 ● The paper **"Body Weight-Length Relationships in Giant Otters" (*Brazilian Archives of Biology and Technology* [2009]: 587–591)** included a scatterplot of y = Weight (in kg) and x = Length (in cm) for a sample of giant otters. Data for 11 of the otters studied (approximated from a graph in the paper) are given in the accompanying table.

Length (cm)	**Weight (kg)**
160	28.9
159	26.3
157	22.1
150	18.7
141	16.5
131	14
124	12.1
115	10.9
85	4.6
66	1.8
68	2.2

Construct a scatterplot for this data set. Would you describe the relationship between length and weight as linear? If not, suggest a transformation that might straighten the plot.

Section 5.5

Interpreting and Communicating the Results of Statistical Analyses

Using either the least-squares line to summarize a linear relationship or a correlation coefficient to describe the strength of a linear relationship is common in investigations that focus on more than a single variable. The methods described in this chapter are among the most widely used statistical tools. When bivariate numerical data are analyzed in journal articles and other published sources, it is common to find a scatterplot of the data and a least-squares line or a correlation coefficient.

Copyright 2025 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s). Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

Communicating the Results of Statistical Analyses

When reporting the results of a data analysis involving bivariate numerical data, it is important to include graphical displays as well as numerical summaries. Including a scatterplot and providing a description of what the plot reveals about the form of the relationship between the two variables under study establishes the context in which numerical summary measures, such as the correlation coefficient or the equation of the least-squares line, can be interpreted.

In general, the goal of an analysis of bivariate numerical data is to learn about the relationship between the two variables. If there is a relationship, you can describe how strong or weak the relationship is. If the goal of the study is to describe the strength of the relationship and the scatterplot shows a linear pattern, the value of the correlation coefficient or the coefficient of determination can be reported as a measure of the strength of the linear relationship.

When interpreting the value of the correlation coefficient, it is a good idea to relate the interpretation to the pattern observed in the scatterplot. This is especially important before making a statement of no relationship between two variables, because a correlation coefficient with a value near 0 does not necessarily imply that there is no relationship of any form. Similarly, a correlation coefficient near 1 or -1, by itself, does not guarantee that the relationship is linear.

If the goal of a study is prediction, then in addition to a scatterplot and the equation of the least-squares line, how well the linear prediction model fits the data should also be addressed. At a minimum, both the values of s_e (the standard deviation about the least-squares line) and r^2 (the coefficient of determination) should be included. Including a residual plot can also provide support for the appropriateness of using a line to describe the relationship between the two variables.

What to Look for in Published Data

Here are a few things to consider when you read an article that includes an analysis of bivariate data:

- What two variables are being studied? Are they both numerical? Is a distinction made between a dependent variable and an independent variable?
- Does the article include a scatterplot of the data? If so, does there appear to be a relationship between the two variables? Can the relationship be described as linear, or is some type of nonlinear relationship a more appropriate description?
- Does the relationship between the two variables appear to be weak or strong? Is the value of a correlation coefficient reported?
- If the least-squares line is used to summarize the relationship between the dependent and independent variables, is there an assessment of how accurate predictions based on the least-squares line might be? For example, are the values of r^2 or s_e reported? How are these values interpreted, and what do they imply about the usefulness of the least-squares line?
- If a correlation coefficient is reported, is it interpreted properly? Be cautious of interpretations that claim a causal relationship.

The authors of the paper **"Recycling and Ambivalence: Quantitative and Qualitative Analyses of Household Recycling Among Young Adults" (*Environment and Behavior* [2008]: 777–797)** describe a study of recycling among young adults in Sweden. They considered y = Recycling behavior (a numerical measure that was based on how often six types of household waste [newspapers, glass, hard plastic, soft plastic, metal, and paper] were recycled) and x = Distance to the nearest recycling facility. They reported that "a check of a plot between the variables recycling behavior and distance to the nearest recycling facility showed an approximately linear association between the two variables." Based on this observation, a least-squares line is an appropriate way to summarize the relationship between recycling behavior and distance to nearest recycling facility.

A related article, **"Rubbish Regression and the Census Undercount" (*Chance* [1992]: 33)**, describes work done by the Garbage Project at the University of Arizona. Project researchers analyzed different categories of garbage for a number of households. They were asked by the Census Bureau to see whether any of the garbage data variables were related to household size. They reported that "the weight data for different categories of garbage were plotted on graphs against data on household size, dwelling by dwelling, and the resulting scatterplots were analyzed to see in which categories the weight showed a steady, monotonic rise relative to household size."

The researchers determined that the strongest linear relationship appeared to be that between the amount of plastic discarded and household size. The line used to summarize this relationship was stated to be $\hat{y} = 0.2815x$, where $y =$ Household size and $x =$ Weight (in pounds) of plastic during a 5-week collection. Note that this line has an intercept of 0. Scientists at the Census Bureau believed that this relationship would extend to entire neighborhoods, and so the amount of plastic discarded by a neighborhood could be measured (rather than having to measure house by house) and then used to approximate the neighborhood size.

An example of the use of the correlation coefficient is found in a paper describing a study of nightingales (a species of bird known for its song). For male songbirds, both physical characteristics and the quality of the song play a role in a female's choice of a mate. The authors of the article **"Song Repertoire Is Correlated with Body Measures and Arrival Date in Common Nightingales" (*Animal Behaviour* [2005]: 211–217)** used data from $n = 20$ nightingales to reach the conclusion that there was a positive correlation between the number of different songs in a nightingale's repertoire and both body weight ($r = 0.53$) and wing length ($r = 0.47$), and a negative correlation between repertoire size and arrival date ($r = -0.47$). This means that heavier birds tend to know more songs, as do birds with longer wings. The authors of the paper indicated that the observed correlation between repertoire size and body characteristics was unlikely to be due solely to the age of the bird, since all nightingales in the study were more than 3 years old and prior research indicates that repertoire size does not continue to increase with age after the third year. The negative correlation between repertoire size and arrival date was interpreted as meaning that male nightingales who knew more songs tended to arrive at their breeding habitats earlier than those who knew fewer songs.

A nonlinear regression was used by the authors of the paper **"Maternal Blood Manganese Levels and Infant Birth Weight" (*Epidemiology* [2009]: 367–373)** to describe the relationship between $y =$ Birth weight and $x =$ Maternal blood-manganese level at delivery for 470 mother-infant pairs. The paper states:

> In this cross-sectional study, there was an inverted U-shaped association between maternal blood-manganese levels at delivery and birth weight in full-term infants. This suggests that both lower and higher manganese exposures are associated with lower birth weight, although the association of higher manganese with lower weight was rather weak and imprecise. This is the first epidemiologic study to provide clear evidence of a nonlinear association between maternal manganese exposure and birth weight.

The paper goes on to suggest why the relationship may be best described by a quadratic rather than a linear equation:

> One possible explanation for this effect would be oxidative stress caused by high manganese levels, leading to impairment of cellular function and growth. Manganese, like iron, is a transitional metal and can catalyze oxidative cellular reactions. Exposure to high levels of iron, a metal with overlapping chemical properties to manganese, has been associated with low birth weight.

A Word to the Wise: Cautions and Limitations

There are a number of ways to get into trouble when analyzing bivariate numerical data! Here are some of the things you need to keep in mind when conducting your own analyses or when reading reports of such analyses:

1. Correlation does not imply causation. A common media blunder is to infer a cause-and-effect relationship between two variables simply because there is a strong

correlation between them. Don't fall into this trap! A strong correlation implies only that the two variables tend to vary together in a predictable way, but there are many possible explanations for why this is occurring besides one variable causing changes in the other.

For example, the article **"Association of Facebook Use With Compromised Well-Being: A Longitudinal Study" (*American Journal of Epidemiology* [2017]: 203–211)** concluded that there was a negative correlation between Facebook use and several measures of well-being based on data from a large sample of people who were thought to be representative of adults in the United States. The negative correlation indicated that for the participants in the study, well-being (which included physical health, mental health, and life satisfaction) tended to decrease as Facebook use increased. But from this correlation alone, it is not reasonable to conclude that Facebook use is the cause of the decrease in well-being. It might be the case that lesser well-being causes increased Facebook use or that a confounding variable, such as loneliness or depression, is related to both Facebook use and well-being. The authors of the paper note, "The associations between Facebook use and compromised well-being may stem from the simple fact that those with compromised well-being attempt to alleviate loneliness by excessively using Facebook."

The paper also looked at the correlation between Facebook use and BMI (body mass index) and found a strong positive correlation between Facebook use and higher BMI. The authors correctly point out that this "suggests that those with higher BMI may be more frequent Facebook users, though it is unlikely that Facebook use itself is the cause of higher BMI."

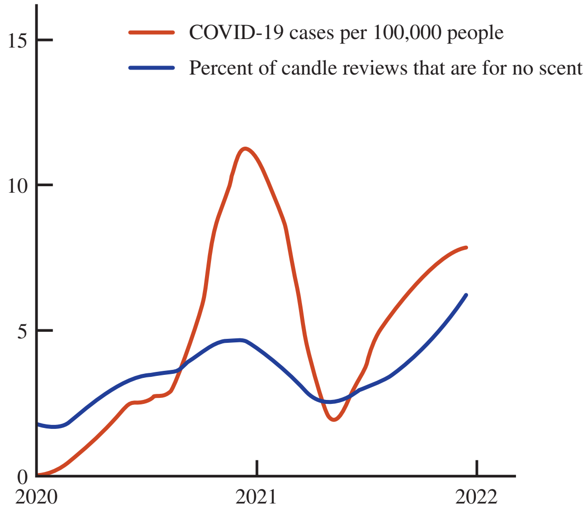
2. A correlation coefficient near 0 does not necessarily imply that there is no relationship between two variables. Before such an interpretation can be given, it is important to examine a scatterplot of the data carefully. Although it may be true that the variables are unrelated, there may in fact be a strong but nonlinear relationship.
3. The least-squares line for predicting y from x is not the same line as the least-squares line for predicting x from y. The least-squares line is, by definition, the line that has the smallest possible sum of squared deviations of points from the line *in the y direction* (it minimizes $\Sigma(y - \hat{y})^2$). The line that minimizes the sum of squared deviations in the y direction is not generally the same as the line that minimizes the sum of the squared deviations in the x direction. So, for example, it is not appropriate to fit a line to data using $y =$ House price and $x =$ House size and then use the resulting least-squares line Price $= a + b$(Size) to predict the size of a house by substituting in a price and then solving for size. Make sure that the dependent and independent variables are clearly identified and that the appropriate line is fit.
4. Beware of extrapolation. It is dangerous to assume that a line used to describe the relationship between x and y is valid over a wider range of x values. Using the least-squares line to make predictions outside the range of x values in the data set often leads to poor predictions.
5. Be careful in interpreting the value of the slope and intercept in the least-squares line. In particular, in many instances interpreting the intercept as the value of y that would be predicted when $x = 0$ is equivalent to extrapolating beyond the range of the x values in the data set. Interpreting the intercept should be avoided unless $x = 0$ is within the range of the data.
6. Remember that the least-squares line may be the "best" line (in that it has a smaller sum of squared deviations than any other line), but that doesn't necessarily mean that the line will produce good predictions. Be cautious of predictions based on a least-squares line without any assessment of the accuracy of predictions, such as s_e and r^2.
7. It is not enough to look at just r^2 or just s_e when assessing a linear model. These two measures address different aspects of the fit of the line. In general, you would like to have a small value for s_e (which indicates that deviations from the line tend to be small) and a large value for r^2 (which indicates that the linear relationship explains a large proportion of the variability in the y values). It is possible to have a small s_e combined with a small r^2 or a large r^2 combined with a large s_e. Remember to consider both values.

Copyright 2025 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s). Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

8. The value of the correlation coefficient as well as the values for the intercept and slope of the least-squares line can be sensitive to influential observations in the data set, particularly if the sample size is small. Because potentially influential observations are those whose x values are far away from most of the x values in the data set, it is important to look for such observations when examining the scatterplot. (Another good reason for *always* starting with a plot of the data!)
9. If the relationship between two variables is not linear, it is preferable to model the relationship using a curve rather than fitting a line to the data. A plot of the residuals from the least-squares line is particularly useful in determining whether a nonlinear model would be a more appropriate choice.

Exercises 5.64–5.67

5.64 The article **"As Mass Testing Ends, How Will Covid Waves Be Tracked?" (*The Economist*, Graphic Detail Daily Chart, December 2, 2022)** describes the work of a researcher at Northeastern University that was published in the *Proceedings of the Sixteenth International AAAI Conference*. This research was motivated by posts on Twitter about negative reviews of scented candles that appeared on Amazon. These negative reviews were posted by unhappy customers who purchased scented candles and complained that the candles had no scent. Some people on Twitter posted screen shots of these negative reviews and suggested that the reviewer probably had COVID-19 since one possible symptom of the virus is loss of the senses of taste and smell. The researcher decided to investigate to see if in fact the number of scented candle complaints might be used as an indication of COVID-19 surges. A graph similar to one that appears in the paper is given in the accompanying figure. This graph shows the number of COVID-19 cases per 100,000 people over time and the percentage of all Amazon reviews of Yankee scented candles that mentioned no smell or no scent over time.



a. Based on the graph, do you think that there is a correlation between the number of COVID-19 cases and the percent of no smell reviews? Would the value of the correlation coefficient be positive or negative?

b. The researcher concluded that the percent of no smell reviews did reflect changes in the number of COVID-19 cases and suggested that social media sources such as Amazon reviews may be useful for identification of rising case numbers. What aspect of the graph supports this conclusion?

5.65 The following quote is from the paper **"Evaluation of the Accuracy of Different Methods Used to Estimate Weights in the Pediatric Population" (*Pediatrics* [2009]: e1045–e1051)**:

> As expected, the model demonstrated that weight increased with age, but visual inspection of an age versus weight plot demonstrated a nonlinear relationship unless infants and children were analyzed separately. The linear coefficient for age as a predictor of weight was 6.93 in infants and 3.1 to 3.48 in children.

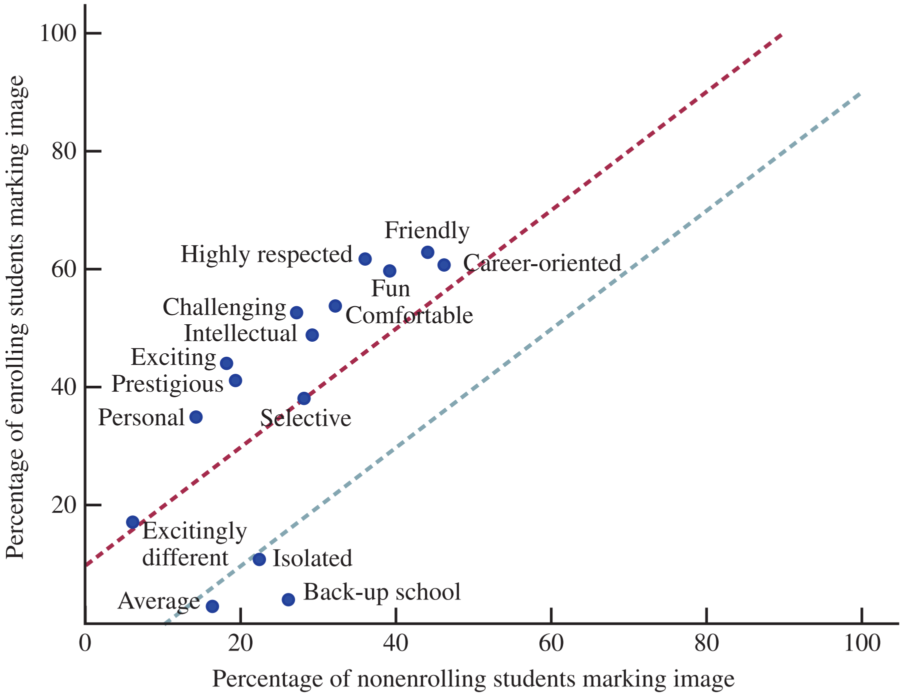
This quote suggests that when a scatterplot of weight versus age was constructed for all 1011 children in the study described in the paper, the relationship between y = weight and x = age was not linear. When the 1011 children were separated into two groups—infants (age birth to 1 year) and children (age 1 to 10 years)—and separate scatterplots were constructed, the relationship between weight and age appeared linear in each scatterplot. The slopes reported in the given quote (referred to as "the linear coefficient") are expressed in kg/year.

Briefly explain why the relationship between weight and age in the scatterplot for the combined group would appear nonlinear.

5.66 The **"Admitted Students Highlights Report 2009"** prepared by **The College Board** for Cal Poly San Luis Obispo summarizes responses to a survey

Copyright 2025 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s). Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

Figure For Exercise 5.66



completed by 2001 new students who enrolled at Cal Poly in fall 2008 and by 2000 students who were admitted to Cal Poly for the fall 2008 term but who enrolled at other universities.

One question in the survey presented a list of "college images" (such as career-oriented and friendly) and asked students to indicate for each image whether or not they associated that image with Cal Poly. The percentage that associated an image with Cal Poly was recorded for enrolling students and for non-enrolling students for each image. For example, 61% of enrolling students but only 46% of non-enrolling students associated the image "career-oriented" with Cal Poly.

The resulting data were used to construct a scatterplot that appeared in the report. The scatterplot is reproduced in the figure at the top of the page.

a. What do you think the two dashed lines in the scatterplot represent?

b. Write a short article appropriate for a student newspaper commenting on what can be learned from this scatterplot. You can assume that the scatterplot will appear with the article.

5.67 The following is an excerpt from a letter to the editor written by Roger Cleary that appeared in the ***San Luis Obispo Tribune* (September 16, 2008)**:

> The causes of poor fuel economy have nothing to do with higher highway speeds, notwithstanding all the press hoopla, including the July 19 Miami Herald claim that "There is no question that slower speeds will save gasoline," and the July 3 statement by Drive Smarter Challenge vehicle director Deron Lovaas that, "I'm not sure whether most people make the connection between how fast they drive and how much fuel they use."
>
> I decided to gather the speed facts for myself using my Chevy, which comes equipped with a fuel usage driver information center, real-time read out. At a road speed of 17.5 mph, it averages 10 mpg; at 35 mph, it averages 20 mpg; and at 65 mph, it averages 30 mpg, all testing done with engine speed standardized at 2000 rpm.
>
> The higher the speed, the better the fuel economy. The faster you drive, the more fuel efficient you become and the more gasoline you save.

Notice that the only speeds that the letter writer provides data for are 17.5, 35, and 65 mph. Studies of the relationship between $y =$ Gas mileage and $x =$ Speed have suggested that the relationship is not linear, and some have used a quadratic curve to describe the relationship between gas mileage and speed.

Write a response to Mr. Cleary that explains how his three observed data points could still be consistent with the statement that higher highway speeds lead to reduced fuel efficiency. Include a graph to support your explanation.

Copyright 2025 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s). Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

Chapter Activities

Activity 5.1 Age and Flexibility

The purpose of this activity is to connect what you learned about collecting data in reasonable ways in Chapter 2 with what you have learned about using a least-squares line to summarize a relationship.

Materials needed: Yardsticks.

In this activity, you will investigate the relationship between age and a measure of flexibility. Flexibility will be measured by asking a person to bend at the waist as far as possible, extending their arms toward the floor. Using a yardstick, the distance from the floor to the fingertip closest to the floor will be measured.

1. Age and the measure of flexibility just described will be measured for a group of individuals. Your goal is to determine whether there is a relationship between age and this measure of flexibility. What are two reasons why it would not be a good idea to use just the students in your class as the subjects for your study?
2. Working as a class, decide on a reasonable way to collect data on the two variables of interest.
3. After your class has collected appropriate data, use the data to construct a scatterplot. Comment on the interesting features of the plot. Does it look like there is a relationship between age and flexibility?
4. If there appears to be a relationship between age and flexibility, fit a model that is appropriate for describing the relationship.
5. *In the context of this activity*, write a brief description of the danger of extrapolation.

Summary Key Concepts and Formulas

Term or Formula	Comment
Scatterplot	A graph of bivariate numerical data in which each observation (x, y) is represented as a point located with respect to a horizontal x axis and a vertical y axis.
Sample correlation coefficient $r = \dfrac{\sum z_x z_y}{n - 1}$	A measure of the extent to which sample x and y values are linearly related. Values close to 1 or -1 indicate a strong linear relationship.
Principle of least squares	The method used to select a line that summarizes an approximate linear relationship between x and y. The least-squares line is the line that minimizes the sum of the squared errors (vertical deviations) for the points in the scatterplot.
$b = \dfrac{\sum(x - \bar{x})(y - \bar{y})}{\sum(x - \bar{x})^2} = \dfrac{\sum xy - \dfrac{(\sum x)(\sum y)}{n}}{\sum x^2 - \dfrac{(\sum x)^2}{n}}$	The slope of the least-squares line.
$a = \bar{y} - b\bar{x}$	The intercept of the least-squares line.
Predicted (fitted) values $\hat{y}_1, \hat{y}_2, \ldots, \hat{y}_n$	Obtained by substituting the x value for each observation in the data set into the least-squares line; $\hat{y}_1 = a + bx_1, \ldots, \hat{y}_n = a + bx_n$
Residuals	Obtained by subtracting each predicted value from the corresponding observed y value: $y_1 - \hat{y}_1, \ldots, y_n - \hat{y}_n$. These are the vertical deviations from the least-squares line.
Residual plot	Scatterplot of the (x, residual) pairs. Isolated points or a pattern of points in a residual plot are indicative of potential problems.
Residual (error) sum of squares $\text{SSResid} = \sum(y - \hat{y})^2$	The sum of the squared residuals is a measure of y variation that cannot be attributed to an approximate linear relationship (unexplained variation).
Total sum of squares $\text{SSTo} = \sum(y - \bar{y})^2$	The sum of squared deviations from the sample mean is a measure of total variation in the observed y values.
Coefficient of determination $r^2 = 1 - \dfrac{\text{SSResid}}{\text{SSTo}}$	The proportion of variation in observed y's that can be explained by an approximate linear relationship.
Standard deviation about the least-squares line $s_e = \sqrt{\dfrac{\text{SSResid}}{n - 2}}$	The size of a "typical" deviation from the least-squares line.
Transformation	A simple function of the x and/or y variable, which is then used in a regression.

Chapter Review Exercises 5.68–5.79

● Data set available online

5.68 ● Is living in a large high-rise apartment building a disadvantage in a medical emergency? This question is investigated in the paper **"Impact of Building Height and Volume on Cardiac Arrest Response Time" (*Prehospital Emergency Care* [2016]: 212–219)**. The accompanying data on the median time (in minutes) it took an emergency team responding to a cardiac arrest call to get from their vehicle at a building entrance to the patient and administer a defibrillator shock (y) and a measure of building size (x, the natural logarithm of the building volume) are approximate values read from a scatterplot that appeared in the paper.

Measure of Building Size	Median Curb-to-Defib Time
8	1.9
9	2
10	2
11	2.3
12	2.5
13	2.6
14	2.9
15	3.1

a. Find the equation of the least-squares regression line.

b. Interpret the slope of the least-squares regression line in the context of this study.

c. Does it make sense to interpret the intercept of the least-squares regression line? If so, give an interpretation. If not, explain why it is not appropriate for this data set.

d. Use the least-squares regression line to predict the median curb-to-defib time for people who live in buildings with a measure of building size of 12.5.

5.69 The authors of the paper **"Flat-Footedness Is Not a Disadvantage for Athletic Performance in Children Aged 11 to 15 Years" (*Pediatrics* [2009]: e386–e392)** studied the relationship between y = Arch height and scores on five different motor ability tests. They reported the following correlation coefficients:

Motor Ability Test	Correlation Between Test Score and Arch Height
Height of counter movement jump	-0.02
Hopping: average height	-0.10
Hopping: average power	-0.09
Balance, closed eyes, one leg	0.04
Toe flexion	0.05

a. Interpret the value of the correlation coefficient between average hopping height and arch height. What does the fact that the correlation coefficient is negative say about the relationship? Do larger arch heights tend to be paired with larger or smaller average hopping heights?

b. The title of the paper suggests that having a small value for arch height (flat-footedness) is not a disadvantage when it comes to motor skills. Do the given correlation coefficients support this conclusion? Explain.

5.70 ● The accompanying data are x = Cost (cents per serving) and y = Fiber content (grams per serving) for 18 high-fiber cereals rated by ***Consumer Reports***.

Cost per Serving	Fiber per Serving	Cost per Serving	Fiber per Serving
33	7	53	13
46	10	53	10
49	10	67	8
62	7	43	12
41	8	48	7
19	7	28	14
77	12	54	7
71	12	27	8
30	8	58	8

a. Construct a scatterplot of y = Fiber content and x = Cost. Based on the scatterplot, how would you describe the relationship between fiber content and cost?

b. Calculate and interpret the value of the correlation coefficient.

c. The serving size differed for the different cereals, with serving sizes varying from a half cup to one and a quarter cups. Converting price and fiber content to "per cup" rather than "per serving" results in the following data. Is the correlation coefficient for the per cup data greater than or less than the correlation coefficient for the per serving data?

Cost per Cup	Fiber per Cup	Cost per Cup	Fiber per Cup
44.0	9.3	53.0	13.0
46.0	10.0	53.0	10.0
49.0	10.0	67.0	8.0
62.0	7.0	43.0	12.0
32.8	6.4	48.0	7.0
19.0	7.0	56.0	28.0
77.0	12.0	54.0	7.0
56.8	9.6	54.0	16.0
30.0	8.0	77.3	10.7

Copyright 2025 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s). Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

5.71 For each of the following pairs of variables, indicate whether you expect a positive correlation, a negative correlation, or a correlation coefficient close to 0. Explain your choice.

a. Price and weight of an apple

b. A person's height and the number of pets they have

c. Time spent studying for an exam and score on the exam

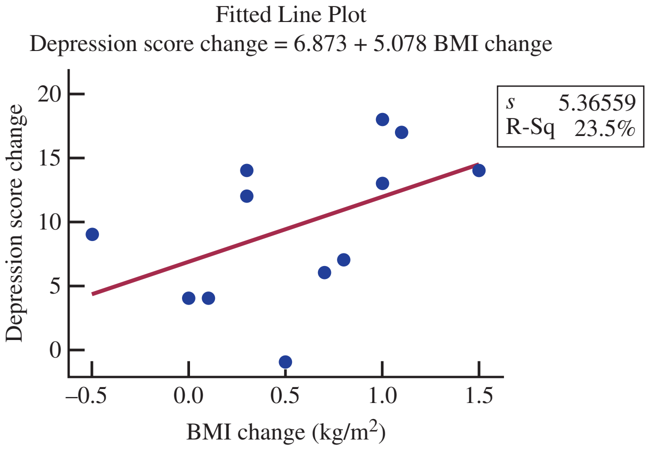
5.72 ● The paper **"Depression, Body Mass Index, and Chronic Obstructive Pulmonary Disease—A Holistic Approach" (*International Journal of COPD* [2016]:239–249)** gave data on change in body mass index (BMI in kilograms/meter^2) and change in a measure of depression for patients suffering from depression who participated in a pulmonary rehabilitation program. The data in the table below are a subset of the data given in the paper and are approximate values read from a scatterplot in the paper.

BMI Change (kg/m^2)	Depression Score Change
0.5	−1
−0.5	9
0.0	4
0.1	4
0.7	6
0.8	7
1.0	13
1.5	14
1.1	17
1.0	18
0.3	12
0.3	14

a. Find the equation of the least-squares line that would allow you to predict change in depression score based on change in BMI.

b. Find the values of r^2 and s_e. Based on these values, do you think that the least-squares line does a good job of describing the relationship between change in depression score and change in BMI? Explain.

c. The following graph is a scatterplot of these data. The least-squares line is also shown. Which observations are outliers? Do the observations with the largest residuals correspond to the patients with the largest change in BMI?



5.73 ● The paper **"Aspects of Food Finding by Wintering Bald Eagles" (*The Auk* [1983]: 477–484)** examined the relationship between the time that eagles spend aerially searching for food (indicated by the percentage of eagles soaring) and relative food availability. The accompanying data were taken from a scatterplot that appeared in this paper. Use x to denote salmon availability and y to denote the percentage of eagles in the air.

x	0	0	0.2	0.5	0.5	1.0
y	28.2	69.0	27.0	38.5	48.4	31.1
x	1.2	1.9	2.6	3.3	4.7	6.5
y	26.9	8.2	4.6	7.4	7.0	6.8

a. Draw a scatterplot for this data set. Would you describe the pattern in the plot as linear or curved?

b. One possible transformation that might lead to a more nearly linear plot involves taking the square root of both the x and y values. Explain why this might be a reasonable transformation.

c. Construct a scatterplot using the variables $\sqrt{x}$ and $\sqrt{y}$. Is this scatterplot more nearly linear than the scatterplot in Part (a)?

d. Suggest another transformation that might be used to straighten the original plot.

5.74 Data on salmon availability (x) and the percentage of eagles in the air (y) were given in the previous exercise.

a. Calculate the correlation coefficient for these data.

b. Because the scatterplot of the original data appeared curved, transforming both the x and y values by taking square roots was suggested. Calculate the correlation coefficient for the variables $\sqrt{x}$ and $\sqrt{y}$. How does this value compare with that calculated in Part (a)? Does this indicate that the transformation was successful in straightening the plot?

Copyright 2025 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s). Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

5.75 ● No tortilla chip lover likes soggy chips, so it is important to find characteristics of the production process that produce chips with an appealing texture. The accompanying data on x = Frying time (in seconds) and y = Moisture content (%) appeared in the paper, **"Thermal and Physical Properties of Tortilla Chips as a Function of Frying Time" (*Journal of Food Processing and Preservation* [1995]: 175–189)**:

Frying time (x):	5	10	15	20	25	30	45	60
Moisture content (y):	16.3	9.7	8.1	4.2	3.4	2.9	1.9	1.3

a. Construct a scatterplot of these data. Does the relationship between moisture content and frying time appear to be linear?

b. Transform the y values using $y' = \log(y)$ and construct a scatterplot of the (x, y') pairs. Does this scatterplot look more nearly linear than the one in Part (a)?

c. Find the equation of the least-squares line that describes the relationship between y' and x.

d. Use the least-squares line from Part (c) to predict moisture content for a frying time of 35 minutes.

5.76 Explain why it can be misleading to use the least squares regression line to obtain predictions for x values that are substantially larger or smaller than the x values in the data set.

5.77 The following quote is from the paper **"The Weight of the Bottle as a Possible Extrinsic Clue with Which to Estimate the Price (and Quality) of the Wine? Observed Correlations" (*Food Quality and Preference* [2012]: 41–45)**:

> The weight of the wine bottles was positively correlated ($r = 0.12$; $p < 0.001$) with the price of the wines

a. Does the value of the correlation coefficient indicate that the relationship is weak, moderate, or strong?

b. What is the value of r^2? Write a sentence that gives an interpretation of this value.

5.78 ● The accompanying data on x = Average energy density (calories per 100 grams) and y = Average cost (in dollars) for eight different food groups are from the paper **"The Cost of U.S. Foods as Related to Their Nutritional Value" (*American Journal of Clinical Nutrition* [2010]: 1181–1188)**. The authors were interested in the relationship between average cost and the average energy density of foods.

Food Group	Average Cost	Average Energy Density
Milk and milk products	0.40	182
Eggs	0.32	171
Dry beans, legumes, nuts, and seeds	0.50	330
Grain products	0.47	337
Fruit	0.28	67
Vegetables	0.33	83
Fats, oils, and salad dressings	0.37	390
Sugars, sweets, and beverages	0.40	242

a. Construct a scatterplot of these data. Does the relationship between cost and energy density look approximately linear?

b. Find the equation of the least-squares regression line.

c. Construct a residual plot and comment on any unusual features of the residual plot.

d. Calculate and interpret the value of r^2 in the context of this study.

e. Explain why it would also be informative to look at the value of s_e in addition to the value of r^2 when evaluating the usefulness of the least squares regression line.

f. Calculate and interpret the value of s_e.

g. Based on your answers to Parts (c)–(f), do you think it is appropriate to use the least squares regression line to predict average cost for a food group based on its average energy density? Explain.

h. The meat, poultry, and fish food group has an average energy density of 196. What average cost would you predict for this food group?

5.79 Consider the four (x, y) pairs $(0, 0)$, $(1, 1)$, $(1, -1)$, and $(2, 0)$.

a. What is the value of the sample correlation coefficient r?

b. If a fifth observation is made at the value $x = 6$, find a value of y for which $r > 0.5$.

c. If a fifth observation is made at the value $x = 6$, find a value of y for which $r < 0.5$.

Copyright 2025 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s). Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

Technology Notes

Correlation

TI-83/84

Note: Before beginning this chapter, press the **2nd** key then the **0** key and scroll down to the entry DiagnosticOn. Press **ENTER** twice. After doing this, the correlation coefficient *r* will appear as output with the least-squares line.

1. Enter the data for the independent variable into **L1** (to access lists press the **STAT** key, highlight the option called **Edit…** then press **ENTER**)
2. Input the data for the dependent variable into **L2**
3. Press the **STAT** key
4. Highlight **CALC** then select **LinReg(a+bx)** and press **ENTER**
5. Press the **2nd** key then the **1** key
6. Press ,
7. Press the **2nd** key then the **2** key
8. Press **ENTER**

TI-Nspire

1. Enter the data for the independent variable into a data list (to access data lists select the spreadsheet option and press **enter**)

   Note: Be sure to title the list by selecting the top row of the column and typing a title.

2. Enter the data for the dependent variable into a separate data list
3. Press the **menu** key and select **4:Statistics** then **1:Stat Calculations** then **4: Linear Regression (a+bx)...** and press **enter**
4. For **X List**: select the column with the independent variable data from the drop-down menu
5. For **Y List:** select the column with the dependent variable data from the drop-down menu
6. Press **OK**

Note: You may need to scroll to view the correlation coefficient in the list of output.

JMP

1. Input the data for the dependent variable into one column
2. Input the data for the independent variable into another column
3. Click **Analyze** then select **Multivariate Methods** then select **Multivariate**
4. Click and drag the column name containing the dependent variable from the box under **Select Columns** to the box next to **Y, Columns**
5. Click and drag the column name containing the independent variable from the box under **Select Columns** to the box next to **Y, Columns**
6. Click **OK**

Note: This produces a table of correlations. The correlation between the two variables can be found in the first row, second column.

Minitab

1. Input the data for the dependent variable into the first column
2. Input the data for the independent variable into the second column
3. Select **Stat** then **Regression** then **Regression** then **Fit Regression Model...**
4. Double-click each column name in order to move it to the box under **Variables:**
5. Click **OK**

SPSS

1. Input the data for the dependent variable into the first column
2. Input the data for the independent variable into the second column
3. Select **Analyze** then choose **Correlate** then choose **Bivariate…**
4. Highlight the name of both columns by holding the **ctrl** key and clicking on each name
5. Click the arrow button to move both variables to the **Variables** box
6. Click **OK**

Note: This produces a table of correlations. The correlation between the two variables can be found in the first row, second column.

Note: The correlation can also be produced by following the steps to produce the regression equation.

Excel

1. Input the data into two separate columns
2. Click the **Data** ribbon and select **Data Analysis**
3. **Note:** If you do not see **Data Analysis** listed on the Ribbon, see the Technology Notes for Chapter 2 for instructions on installing this add-on.
4. Select **Correlation** from the dialog box and click **OK**
5. Click in the box next to **Input Range:** and select BOTH columns of data (if you input and selected titles for the columns, click the box next to **Labels in First Row**)
6. Click **OK**

Note: The correlation between the variables can be found in the first column, second row of the table that is output.

Regression

TI-83/84

1. Enter the data for the independent variable into **L1** (to access lists press the **STAT** key, highlight the option called **Edit…** then press **ENTER**)
2. Input the data for the dependent variable into **L2**
3. Press the **STAT** key
4. Highlight **CALC** then select **LinReg(a+bx)** and press **ENTER**
5. Press the **2nd** key then the **1** key
6. Press ,

7. Press the **2nd** key then the **2** key
8. Press **ENTER**

TI-Nspire

1. Enter the data for the independent variable into a data list (to access data lists select the spreadsheet option and press **enter**)

   Note: Be sure to title the list by selecting the top row of the column and typing a title.

2. Enter the data for the dependent variable into a separate data list
3. Press the **menu** key and select **4:Statistics** then **1:Stat Calculations** then **4:Linear Regression (a+bx)...** and press **enter**
4. For **X List:** select the column with the independent variable data from the drop-down menu
5. For **Y List:** select the column with the dependent variable data from the drop-down menu
6. Press **OK**

JMP

1. Enter the data for the dependent variable into the first column
2. Input the data for the independent variable into the second column
3. Click **Analyze** then select **Fit Y by X**
4. Click and drag the column name containing the dependent data from the box under **Select Columns** to the box next to **Y, Response**
5. Click and drag the column name containing the independent data from the box under **Select Columns** to the box next to **X, Factor**
6. Click **OK**
7. Click the red arrow next to **Bivariate Fit…**
8. Click **Fit Line**

Minitab

1. Input the data for the dependent variable into the first column
2. Input the data for the independent variable into the second column
3. Select **Stat** then **Regression** then **Regression** then **Fit Regression Model...**
4. Highlight the name of the column containing the dependent variable and click **Select** to add it to the **Responses:** box
5. Click in the **Continuous predictors:** box then highlight the name of the column containing the independent variable and click **Select** to add it to this box
6. Click **OK**

Note: You may need to scroll up in the Session window to view the regression equation.

SPSS

1. Input the data for the dependent variable into the first column
2. Input the data for the independent variable into the second column
3. Select **Analyze** then choose **Regression** then choose **Linear…**
4. Highlight the name of the column containing the dependent variable
5. Click the arrow button next to the **Dependent** box to move the variable to this box
6. Highlight the name of the column containing the independent variable
7. Click the arrow button next to the **Independent** box to move the variable to this box
8. Click **OK**

Note: The regression coefficients can be found in the **Coefficients** table. The intercept value can be found in the first column of the (Constant) row. The value of the slope can be found in the first column of the row labeled with the independent variable name.

Excel

1. Input the data into two separate columns
2. Click the **Data** ribbon and select **Data Analysis**
3. **Note:** If you do not see Data Analysis listed on the Ribbon, see the Technology Notes for Chapter 2 for instructions on installing this add-on.
4. Select **Regression** from the dialog box and click **OK**
5. Click in the box next to **Input Y Range:** and select the dependent variable data
6. Click in the box next to **Input X Range:** and select the independent variable data (if you input and selected titles for BOTH columns, check the box next to **Labels**)
7. Click **OK**

Note: The regression coefficients can be found in the third table under the **Coefficients** column.

Residuals

TI 83/84

Note: When the TI 83/84 performs a regression analysis, the residuals are automatically stored in a list called "RESID." To prevent loss of the residuals, copy them into a named list.

1. Press the **2nd** key then the List key. **Select RESID**. (You may have to scroll down to find it.) Your screen should now show "ʟRESID."
2. Click on **STO**, then the **2nd** key, and then **L3**.
3. Click on Enter.

TI-Nspire

1. Enter the data for the independent variable into a data list (to access data lists select the spreadsheet option and press **enter**)

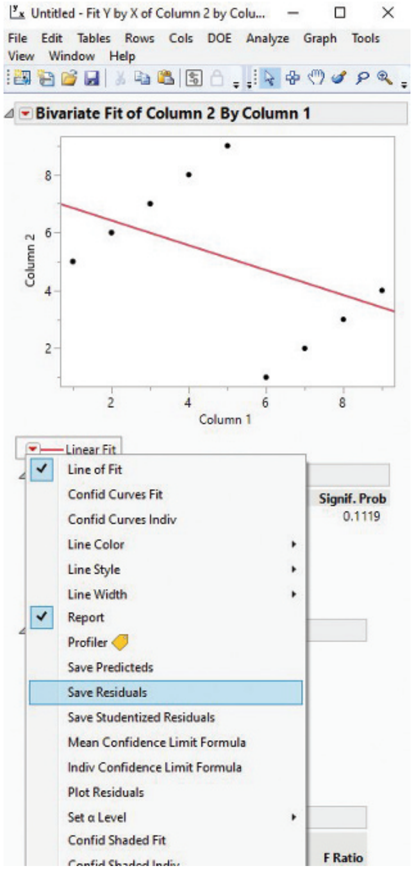
   Note: Be sure to title the list by selecting the top row of the column and typing a title.

2. Enter the data for the dependent variable into a separate data list
3. Press the menu key and select **4:Statistics** then **1:Stat Calculations** then **4:Linear Regression(a+bx)…** and press **enter**
4. For **X List:** select the column with the independent variable data from the drop-down menu
5. For **Y List:** select the column with the dependent variable data from the drop-down menu
6. Press **OK**

Copyright 2025 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s). Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

JMP

1. Enter the data for the dependent variable into the first column
2. Input the data for the independent variable into the second column
3. Click **Analyze** then select **Fit Y by X**
4. Click and drag the column name containing the dependent data from the box under **Select Columns** to the box next to **Y, Response**
5. Click and drag the column name containing the independent data from the box under **Select Columns** to the box next to **X, Factor**
6. Click **OK**
7. Click the red arrow next to **Bivariate Fit…**
8. Click **Fit Line**
9. Click the red arrow next to **Linear Fit**
10. Click **Save Residuals**


Untitled - Fit Y by X of Column 2 by Colu...
File Edit Tables Rows Cols DOE Analyze Graph Tools View Window Help
Bivariate Fit of Column 2 By Column 1
Column 2
8
6
4
2
2
4
6
8
Column 1
Linear Fit
Line of Fit
Confid Curves Fit
Confid Curves Indiv
Line Color
Line Style
Line Width
Report
Profiler
Save Predicteds
Save Residuals
Save Studentized Residuals
Mean Confidence Limit Formula
Indiv Confidence Limit Formula
Plot Residuals
Set α Level
Confid Shaded Fit
Confid Shaded Indiv
Signif. Prob
0.1119
F Ratio

Source: JMP

Source: JMP

Source: Microsoft Corporation

Minitab

1. Input the data for the dependent variable into the first column
2. Input the data for the independent variable into the second column
3. Select **Stat** then **Regression** then **Regression** then **Fit Regression Model...**
4. Highlight the name of the column containing the dependent variable and click **Select** to add it to the **Responses:** box
5. Click in the **Continuous predictors:** box then highlight the name of the column containing the independent variable and click **Select** to add it to this box
6. Click **Storage…**
7. Check the box next to **Residuals**
8. Click **OK**
9. Click **OK**

Note: The residuals will be saved in the Minitab Worksheet in a new column.

SPSS

1. Input the data for the dependent variable into the first column
2. Input the data for the independent variable into the second column
3. Select **Analyze** then choose **Regression** then choose **Linear…**
4. Highlight the name of the column containing the dependent variable
5. Click the arrow button next to the **Dependent** box to move the variable to this box
6. Highlight the name of the column containing the independent variable
7. Click the arrow button next to the **Independent** box to move the variable to this box
8. Click on the **Save…** button
9. Click the check box next to **Unstandardized** under the **Residuals** section
10. Click **Continue**
11. Click **OK**

Note: The residuals will be saved in the SPSS worksheet in a new column.

Excel

1. Input the data into two separate columns
2. Click the **Data** ribbon and select **Data Analysis**

   Note: If you do not see Data Analysis listed on the Ribbon, see the Technology Notes for Chapter 2 for instructions on installing this add-on.

3. Select **Regression** from the dialog box and click **OK**
4. Click in the box next to **Input Y Range:** and select the dependent variable data
5. Click in the box next to **Input X Range:** and select the independent variable data (if you input and selected titles for BOTH columns, check the box next to **Labels**)
6. Check the box next to **Residuals** under the **Residuals** section of the dialog box
7. Click **OK**

Residual Plot

TI-83/84

The TI-83/84 does not have the functionality to produce a residual plot automatically. After using Linreg(a+bx), select 2nd, Statplot. Set the first plot as a scatter with XList: L1 and YList: RESID. Select Zoom, 9: ZoomStat and a residual plot is displayed.

TI-Nspire

The TI-Nspire does not have the functionality to produce a residual plot automatically.

Copyright 2025 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s). Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

JMP

1. Begin by saving the residuals as described in the previous section
2. Form a scatterplot of the independent variable versus the residuals using the procedures described in Chapter 2 for scatterplots

Minitab

1. Input the data for the dependent variable into the first column
2. Input the data for the independent variable into the second column
3. Select **Stat** then **Regression** then **Regression** then **Fit Regression Model...**
4. Highlight the name of the column containing the dependent variable and click **Select**
5. Click in the **Continuous predictors:** box then highlight the name of the column containing the independent variable and click **Select** to add it to this box
6. Click **Graphs…**
7. Click the box under **Residuals versus the variables**:
8. Double-click the name of the independent variable
9. Click **OK**
10. Click **OK**

SPSS

1. Begin by saving the residuals as described in the previous section
2. Form a scatterplot of the independent variable versus the residuals using the procedures described in Chapter 2 for scatterplots

Excel

1. Input the data into two separate columns
2. Click the **Data** ribbon and select **Data Analysis**

Note: If you do not see **Data Analysis** listed on the Ribbon, see the Technology Notes for Chapter 2 for instructions on installing this add-on.

3. Select **Regression** from the dialog box and click **OK**
4. Click in the box next to **Y:** and select the dependent variable data
5. Click in the box next to **X:** and select the independent variable data (if you input and selected titles for BOTH columns, check the box next to **Labels**)
6. Check the box next to **Residuals Plots** under the **Residuals** section of the dialog box
7. Click **OK**

Transforming Variables

TI 83/84

1. Enter the data for the variable you wish to transform into **L1** (to access lists press the **STAT** key, highlight the option called Edit…then press **ENTER**)
2. Click on **Stat** and select **Edit**
3. Using the arrow keys (directly above the **VARS** and **CLEAR** Keys) position the cursor at the top of **List2**. That is, the cursor should be covering the "**L2**." In addition, "**L2=**" should appear in the bottom line

Now choose the transformation you wish to make. For example, to set the values of **L2** equal to the natural log of **L1**, do the following:

4. Click the **LN** button; the formula at the bottom of the screen will change to "**L2=ln(**"
5. Press the **2nd** key then the **1** key
6. Click on the right parenthesis key; the formula at the bottom of the screen will change to "**L2=ln(L1)**"
7. Click on **Enter**

JMP

1. Enter the pre-transformed data into **Column1**
2. Double-click at the top of **Column2**. A window with properties for **Column2** will pop up
3. In that window, click on **Column Properties**, select **Formula**, and click on **Edit Formula**

Now choose the transformation you wish to make. For example, to set the values of Column2 equal to the natural logarithm of Column1, do the following:

4. Click on Transcendental, then click on **Ln**; the formula box will be updated to "**Ln(   )**"
5. In the TableColumns panel, click on **Column1**; the formula box will be updated to "**Ln(Column 1)**"
6. Click **Ok** and close the column properties window
7. Click **OK**

Minitab

1. Enter the pre-transformed data into column "**C1**"
2. Click on Calc in the main menu, and then Calculator
3. Type **C2** in the "Store result in variable" box

Now choose the transformation you wish to make. For example, to set the values of **C2** equal to the natural log of **C1**:

4. Click on Natural log (log bas e) in the Functions drop-down menu and click on **Select**. The "Expression:" window will be updated to "**LN(number)**"
5. Double-click on **C1** in the variables list panel; the "Expression:" window will be updated to "**LN(C1)**"
6. Click **OK**

Automatic Nonlinear Regression

TI 83/84

1. Enter the data for the independent variable into **L1** (to access lists press the **STAT** key, highlight the option called **Edit…** then press **ENTER**)
2. Input the data for the dependent variable into **L2**
3. Press the **STAT** key
4. Select **CALC**

Now choose the nonlinear option you wish; for example, to perform exponential regression, do the following:

5. Select **ExpReg** and press **ENTER**
6. Press the **2nd** key then the **1** key
7. Press,
8. Press the **2nd** key then the **2** key
9. Press **ENTER**

Cumulative Review Exercises CR5.1–CR5.15

● Data set available online

CR5.1 The article **"Rocker Shoe Put to the Test: Can it Really Walk the Walk as a Way to Get in Shape?" (*USA TODAY,* October 12, 2009)** describes claims made by Skechers about Shape-Ups, a shoe line introduced in 2009. These curved-sole sneakers are supposed to help you "get into shape without going to the gym" according to a Skechers advertisement. Briefly describe how you might design a study to investigate this claim. Include how you would select subjects and what variables you would measure. Is the study you designed an observational study or an experiment?

CR5.2 ● Just how dirty is the money you carry in your wallet? The authors of the article **"Dirty Money: New Research Reveals the Filthiness of Our Cash, Cards, & Coins" (lendedu.com, 2020)** measured the bacteria on the surface of credit and debit cards, bills, and coins. Germ scores (with higher scores corresponding to dirtier surfaces) for 27 bills of various denominations ranging from \$1 to \$100 are shown in the accompanying table.

32	35	39	41	54	55	56	68	74
78	98	104	105	108	119	119	126	129
143	163	200	242	259	273	349	609	633

a. Construct a histogram of the germ scores.

b. Based on the histogram, would you expect the mean germ score to be greater than, about equal to, or less than the median germ score?

c. For reference, it is recommended that a food establishment surface should have a germ score of 10 or less. The mean germ scores for other surfaces studied were 68 for a New York City subway pole, 163 for bathrooms at Penn Station in New York City, and 664 for door handles at McDonald's. Given this information, write a few sentences commenting on the germs scores for bills.

CR5.3 ● The article referenced in the previous exercise also gave the germs scores below for 12 coins.

25	32	33	52	60	84
92	109	227	228	329	356

a. Using the data given here and in Exercise CR5.2, construct comparative boxplots that would allow you to compare germs scores of bills and coins.

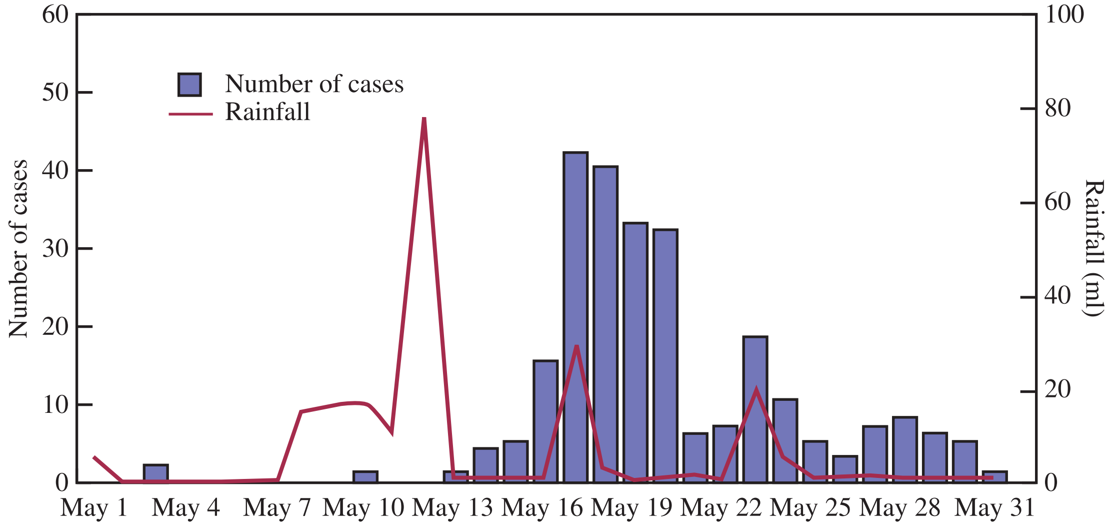
b. Write a few sentences comparing the germ score distributions for bills and coins.

CR5.4 The graph in the figure at the bottom of the page appeared in the report **"Testing the Waters 2009" (Natural Resources Defense Council)**. Spend a few minutes looking at the graph and reading the caption that appears with the graph. Briefly explain how the graph supports the claim that discharges of polluted storm water may be responsible for increased illness levels.

CR5.5 ● Each year the Harris Poll surveys Americans on a number of issues. It uses responses to several questions to calculate a "Happiness" index that measures overall happiness. The article **"Latest Happiness Index Reveals American Happiness at All-Time Low," theharrispoll.com)** included the happiness index for the 7 years between 2008 and 2016. Also

Figure for Exercise CR5.4

Influence of heavy rainfall on occurrence of *E. coli* infections.



The graph shows the relationship between unusually heavy rainfall and the number of confirmed cases of *E. coli* infection that occurred during a massive disease outbreak in Ontario, Quebec, in May 2000. The incubation period for *E. coli* is usually 3 to 4 days, which is consistent with the lag between extreme precipitation events and surges in the number of cases.

included in the article were the percentages of people who responded "somewhat agree" or "strongly agree" to the following statements:

> Statement 1 (happy with life) 1: At this time, I'm generally happy with my life.
>
> Statement 2 (won't benefit): I won't get much benefit from the things that I do anytime soon.

Year	Happiness Index	Happy with Life Statement (percentage somewhat or strongly agree)	Won't Benefit Statement (percentage somewhat or strongly agree)
2008	35	83	32
2009	35	81	38
2010	33	80	36
2011	33	80	38
2013	33	77	42
2015	34	82	36
2016	31	81	41

a. Calculate the value of the correlation coefficient for Happiness index and the response to the Happy with life statement.

b. Calculate the value of the correlation coefficient for Happiness index and the response to the Won't benefit statement.

c. Is there a stronger relationship between Happiness index and the response to Statement 1 or between Happiness index and the response to Statement 2?

d. Write a few sentences describing the relationships between Happiness index and the responses to the two statements.

CR5.6 ● The amount of money spent each year on science, space, and technology in the United States (in millions of dollars) and the amount of money spent on pets in the United States (in billions of dollars) for the years 2000 to 2009 were used to construct the graph at the bottom of the page. (The data are from the website **tylervigen.com/spurious-correlations**).

Based on these time series plots, would the correlation coefficient between amount spent on science, space, and technology and the amount spent on pets be positive or negative? Weak or strong? What aspect of the time series plots support your answer?

CR5.7 Below are the data used to construct the time series plots in the previous exercise. Calculate the value of the correlation coefficient for the amount spent on science, space, and technology and the amount spent on pets. Explain how this value is consistent with your answer to the previous exercise.

Year	Amount Spent on Science, Space, and Technology (millions of dollars)	Amount Spent on Pets (billions of dollars)
2000	18,594	39.7
2001	19,753	41.9
2002	20,734	44.6
2003	20,831	48.8
2004	23,029	49.8
2005	23,597	53.1
2006	23,584	56.9
2007	25,525	61.8
2008	27,731	65.7
2009	29,449	67.1

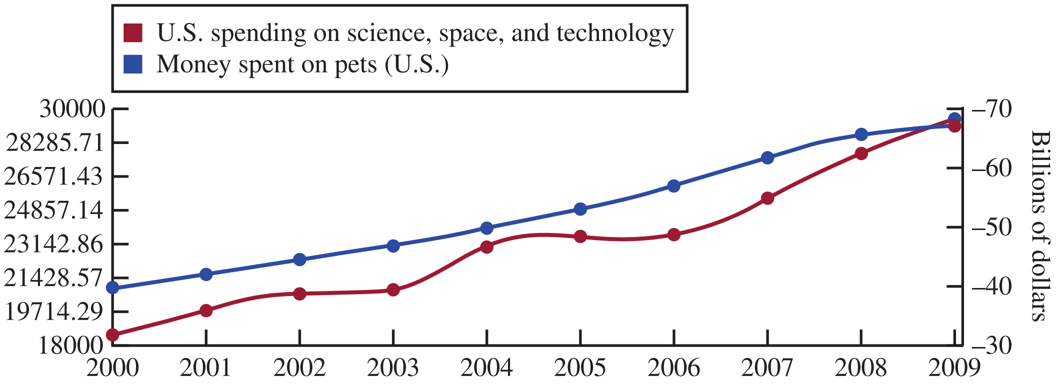
CR5.8 The authors of the paper **"Geographic Access to Pediatric Cancer Care in the US" (*JAMA Network Open*, January 2023)** collected data on how long children with cancer had to travel to receive care for childhood cancers. Based on large samples of children with cancer in each of four geographic regions, they reported the following medians and interquartile ranges.

Region	Median Travel Time (minutes)	IQR
Northeast	11	21
Midwest	22	40
South	24	34
West	17	29

Write a few sentences describing regional differences in the times that children with cancer must travel to receive treatment.

CR5.9 ● The paper **"Total Diet Study Statistics on Element Results" (Food and Drug Administration, April 25, 2000)**

Figure For Exercise CR5.6



gave information on sodium content for various types of foods. Twenty-six tomato catsups were analyzed. Data consistent with summary quantities given in the paper were

Sodium content (mg/kg)

12,148 10,426 10,912 9116 13,226 11,663
11,781 10,680 8457 10,788 12,605 10,591
11,040 10,815 12,962 11,644 10,047 10,478
10,108 12,353 11,778 11,092 11,673 8758
11,145 11,495

Calculate the values of the quartiles and the interquartile range.

CR5.10 ● The paper referenced in Exercise CR5.9 also gave data on sodium content (in milligrams per kilogram) of 10 chocolate puddings made from instant mix:

3099 3112 2401 2824 2682 2510 2297
3959 3068 3700

a. Calculate the mean, the standard deviation, and the interquartile range for sodium content of these chocolate puddings.
b. Based on the interquartile range, is there more or less variability in sodium content for the chocolate pudding data than for the tomato catsup data of Exercise CR5.9?

CR5.11 ● A report from Texas Transportation Institute **(2021 Urban Mobility Report)** reported the average extra travel time (due to traffic congestion) for commute travel per traveler per year in hours for different urban areas. Below are the data for the 15 largest urban areas.

Urban Area	Extra Hours per Traveler per Year
New York	56
Boston	50
Houston	49
San Francisco	46
Washington, DC	42
Los Angeles	41
Dallas–Fort Worth	40
Chicago	39
Atlanta	37
Philadelphia	37
Detroit	35
Seattle	31
Miami	27
Phoenix	25
San Diego	24

a. Calculate the mean and median values for extra travel hours. Based on the values of the mean and median, is the distribution of extra travel hours likely to be approximately symmetric, positively skewed, or negatively skewed?
b. Construct a boxplot that shows outliers for these data and comment on any interesting features of the plot.

CR5.12 ● Cost-to-charge ratios (the percentage of the amount billed that represents the actual cost) for 11 Oregon hospitals of similar size were reported separately for inpatient and outpatient services. The data are shown in the following table.

	Cost-to-Charge Ratio	
Hospital	Inpatient	Outpatient
Blue Mountain	80	62
Curry General	76	66
Good Shepherd	75	63
Grande Ronde	62	51
Harney District	100	54
Lake District	100	75
Pioneer	88	65
St. Anthony	64	56
St. Elizabeth	50	45
Tillamook	54	48
Wallowa Memorial	83	71

a. Does there appear to be a strong linear relationship between the cost-to-charge ratio for inpatient and outpatient services? Justify your answer based on the value of the correlation coefficient and examination of a scatterplot of the data.
b. Are any unusual features of the data evident in the scatterplot?
c. Suppose that the observation for Harney District was removed from the data set. Would the correlation coefficient for the new data set be greater than or less than the one computed in Part (a)? Explain.

CR5.13 In the article **"Reproductive Biology of the Aquatic Salamander *Amphiuma tridactylum* in Louisiana" (*Journal of Herpetology* [1999]: 100–105)**, 14 female salamanders were studied. Using regression, the researchers predicted y = Clutch size (number of salamander eggs) from x = Snout-vent length (in centimeters) as follows:

$$\hat{y} = -147 + 6.175x$$

For the salamanders in the study, the range of snout-vent lengths was approximately 30 to 70 cm.

a. What is the value of the y intercept of the least-squares line?
b. What is the value of the slope of the least-squares line? Interpret the slope in the context of this problem.
c. Would you be reluctant to predict the clutch size when snout-vent length is 22 cm? Explain.

CR5.14 The previous exercise gave the least-squares line for predicting y = Clutch size from x = Snout-vent length. The paper also reported $r^2 = 0.7664$ and SSTo = 43,951.

a. Interpret the value of r^2.
b. Find and interpret the value of s_e (the sample size was $n = 14$).

Copyright 2025 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s). Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

Chapter 6

Probability

Prostock-studio/Shutterstock.com

We make decisions based on uncertainty every day. Should you buy an extended warranty for your new laptop? It depends on the likelihood that it will fail during the warranty period. Should you allow 45 minutes to get to your 8 a.m. class, or is 35 minutes enough? From experience, you may know that most mornings you can drive to school and park in 25 minutes or less. Most of the time, the walk from your parking space to class is 5 minutes or less. But how often will the drive to school or the walk to class take longer than you expect? When it takes longer than usual to drive to campus, is it more likely that it will also take longer to walk to class? less likely? Or are the driving and walking times unrelated?

Some questions involving uncertainty are more serious: Suppose a hospital that needs power to keep life supporting equipment running has a traditional power source and a backup generator. How likely is it that the traditional power source will fail? How likely is it that both the traditional power source and the backup generator will fail? Is one backup generator enough to ensure that the risk of total power loss is small, or is a second backup generator needed? Probability, the systematic study of uncertainty, can be used to answer questions like these.

Learning Objectives

After completing this chapter, you will be able to:

- Interpret a probability in context as a long-run relative frequency.
- Identify the sample space for a chance experiment.
- Explain what it means for two events to be mutually exclusive.
- Explain what it means for two events to be independent.
- Calculate the probability of events, including events that are the union of, the intersection of, or the complement of other events.
- Explain the difference between an unconditional and a conditional probability.
- Calculate conditional probabilities.
- Estimate probabilities empirically and using simulation.

Section 6.1 Chance Experiments and Events

The basic ideas and terminology of probability are most easily introduced in situations that are both familiar and reasonably simple. Because of this, some of the initial examples involve easily explained activities such as tossing a coin, selecting cards from a deck, and rolling a die. However, after considering a few of these simple examples, more interesting and realistic situations will be explored.

Chance Experiments

When a single coin is tossed, there are two possible outcomes. The coin can land with its heads side up or its tails side up. The selection of a single card from a well-mixed standard deck of 52 playing cards may result in the ace of spades, the five of diamonds, or any one of the other 50 possibilities. Situations such as these are referred to as **chance experiments**.

Definition

Chance experiment: The process of making an observation when there is uncertainty about which of two or more possible outcomes will result.

When the term chance experiment is used in a probability setting, it means something different from what was meant by the term experiment in Chapter 2 (where experiments were a type of statistical study that investigates the effect of two or more treatments on a response). For example, in an opinion poll or survey, there is uncertainty about whether an individual selected at random from the population of interest supports a ballot initiative, and when a die is rolled, there is uncertainty about which face will land on the top face. Both of these situations fit the definition of a *chance* experiment.

Consider a chance experiment to investigate whether young adults (age 18–39) or older adults (age 40+) are more likely to choose a hybrid car over a traditional internal combustion engine car when purchasing a Honda Civic at a particular car dealership. The Honda Civic is available with either a hybrid or a traditional engine. In this chance experiment, a customer will be selected at random from those who purchased a Honda Civic. The type of vehicle purchased (hybrid or traditional) will be determined and the customer's age group (young adult or older adult) will be recorded. Before the customer is selected the outcome of this chance experiment is unknown. However, you do know what the possible outcomes are. This set of possible outcomes is called the **sample space**.

Definition

Sample space: The collection of all possible outcomes of a chance experiment.

The sample space of a chance experiment can be represented in many ways. One representation is a simple list of all the possible outcomes. For the car-purchase chance experiment described above, the possible outcomes are:

1. A young adult who bought a hybrid engine
2. An older adult who bought a hybrid engine
3. A young adult who bought a traditional engine
4. An older adult who bought a traditional engine

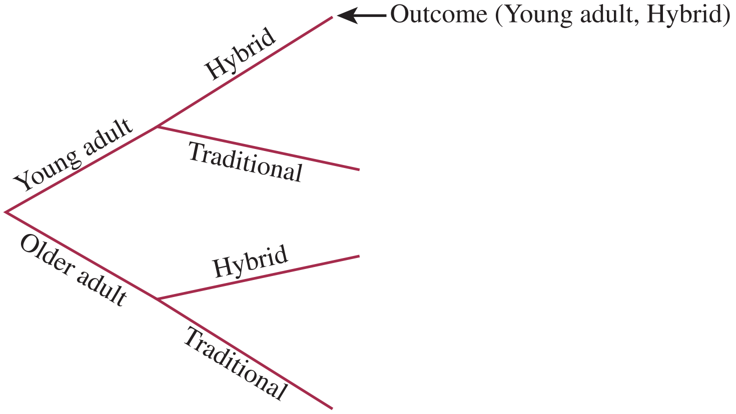
Set notation and ordered pairs can also be used. A young adult who bought a hybrid engine is represented as (young adult, hybrid). The sample space is then

$$\text{sample space} = \left\{\begin{matrix}\text{(young adult, hybrid), (older adult, hybrid),} \\ \text{(young adult, traditional), (older adult, traditional)}\end{matrix}\right\}$$

Copyright 2025 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s). Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

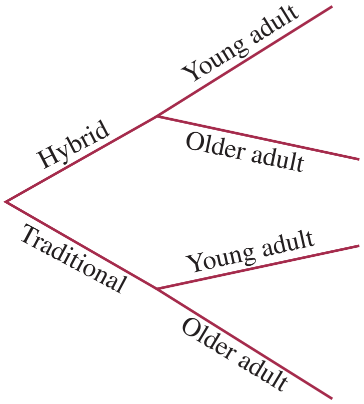
Another useful representation of the sample space is a tree diagram. A tree diagram (shown in Figure 6.1) for the outcomes of the car-purchase chance experiment has two sets of branches corresponding to the two pieces of information that were gathered. To identify any particular outcome in the sample space, traverse the tree by first selecting a branch corresponding to the age group of the selected customer and then a branch identified with an engine type.

Figure 6.1
Tree diagram for the car purchase example.



In the tree diagram of Figure 6.1, there is no particular reason for having the age group of the customer as the first-generation branch and the type of engine as the second-generation branch. Some chance experiments involve making observations in a particular order, in which case the order of the branches in the tree does matter. In this example, however, it would be acceptable to represent the sample space with the tree diagram shown in Figure 6.2.

Figure 6.2
Another tree diagram for the car purchase example.



The sample space for a chance experiment can be represented in several ways, but the representations all have one thing in common: *Every* possible outcome of the chance experiment is included in the representation.

Events

In the car-purchase chance experiment, you might be interested in which particular outcome will result. Or you might focus on a group of outcomes that involve the purchase of a hybrid—the group consisting of (young adult, hybrid) and (older adult, hybrid). When you form a collection of one or more individual outcomes, you are creating what is known as an *event*.

Definitions

Event: Any collection of outcomes from the sample space of a chance experiment.

Simple event: An event consisting of exactly one outcome.

Events are usually represented by an uppercase letter, such as A, B, C, and so on. Sometimes the same letter with different numerical subscripts, such as E_1, E_2, E_3, … may also be used to represent events.

Example 6.1 Car Preferences

Reconsider the situation in which a person who purchased a Honda Civic was categorized by age group (Y for young adult or O for older adult) and type of engine purchased (H = hybrid, T = traditional). Using this notation, one possible representation of the sample space is

sample space = {YH, OH, YT, OT}

Because there are four outcomes, there are four simple events:

E_1 = YH E_2 = OH E_3 = YT E_4 = OT

One event of interest might be the event consisting of all outcomes for which a hybrid was purchased. The event *hybrid* is

hybrid = {YH, OH}

Another event is the event that the purchaser is a young adult,

young adult = {YH, YT}

Example 6.2 Video Game

Suppose you believe that after losing to an opponent in a video game, a player is more likely to lose the next game. You carry out a chance experiment that consists of watching two consecutive games for a particular player and observing whether the player won, tied, or lost each of the two games. In this case (using W, T, and L to represent win, tie, and loss, respectively), the sample space can be represented as

sample space = {WW, WT, WL, TW, TT, TL, LW, LT, LL}

The event *player loses exactly one of the two games*, denoted by L_1, can then be defined as

L_1 = {WL, TL, LW, LT}

Only one outcome (one simple event) occurs when a chance experiment is performed. You would say that a given event occurs whenever one of the outcomes making up the event occurs. For example, if the outcome in the car purchase example is YH, then the simple event *young adult* who bought *a hybrid* has occurred, and so has the event *hybrid purchased*. The event *hybrid purchased* is not a simple event because it is made up of more than one possible outcome.

Forming New Events

Once some events have been specified, there are several ways they can be used to create new events.

Definitions

Let A and B denote two events.

Not A: The event that consists of all outcomes in the sample space that are not in event A. *Not A* is called the **complement** of A and is also sometimes denoted by $\boldsymbol{A^C, A', \text{or } \bar{A}}$.

A or B: The event that consists of all outcomes in the sample space that are in at least one of the two events (in A or in B or in both). *A or B* is called the **union** of the two events and is also sometimes denoted by $\boldsymbol{A \cup B}$.

A and B: The event that consists of all outcomes in the sample space that are in both the event A and the event B. *A and B* is called the **intersection** of the two events and is also sometimes denoted by $\boldsymbol{A \cap B}$.

Copyright 2025 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s). Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

Example 6.3 Turning Directions

A traffic engineer has been asked to consider whether a stop sign at the bottom of a freeway off-ramp should be replaced by a traffic light. To help in this decision, they plan to observe traffic patterns for this off-ramp. Suppose they were to record the turning direction (L = left or R = right) of each of three successive vehicles. This is a chance experiment and the sample space contains eight outcomes:

all 3 cars turn left ↘ ↙ *the 1st car turns left and the next 2 turn right*

{LLL, RLL, LRL, LLR, RRL, RLR, LRR, RRR}

Each of these outcomes determines a simple event. Other events include

A = event that exactly one of the cars turns right = {RLL, LRL, LLR}
B = event that at most one of the cars turns right = {LLL, RLL, LRL, LLR}
C = event that all cars turn in the same direction = {LLL, RRR}

Some other events that can be formed from the events just defined are

not $C = C^C$ = event that not all cars turn in the same direction
= {RLL, LRL, LLR, RRL, RLR, LRR}

A *or* $C = A \cup C$ = event that exactly one of the cars turns right or all cars turn in the same direction
= {RLL, LRL, LLR, LLL, RRR}

B *and* $C = B \cap C$ = event that at most one car turns right and all cars turn in the same direction
= {LLL}

Example 6.4 More on Video Games

In Example 6.2, in addition to the event that a video game player loses exactly one of the two games, L_1 = {WL, TL, LW, LT}, you could also define events corresponding to L_0 = neither game is lost and L_2 = both games are lost. Then L_0 = {WW, WT, TW, TT} and L_2 = {LL}.

Some other events that can be formed from those just defined include

L_2^C = event that at most one game was lost
= {WW, WT, WL, TW, TT, TL, LW, LT}

$L_1 \cup L_2$ = event that at least one game was lost
= {WL, TL, LW, LT, LL}

$L_1 \cap L_2$ = event that exactly one game was lost and two games were lost
= the empty set (there are no outcomes that are in both L_1 and L_2)

Sometimes two events have no common outcomes, as was the case for events L_1 and L_2 in the video game example. Such events are described using special terminology.

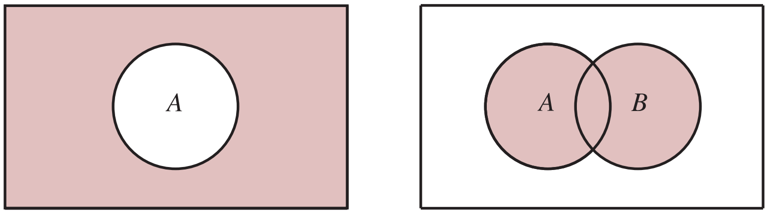
Definition

Mutually exclusive: Two events are **mutually exclusive** if they have no outcomes in common. The term **disjoint** is also sometimes used to describe events that have no outcomes in common.

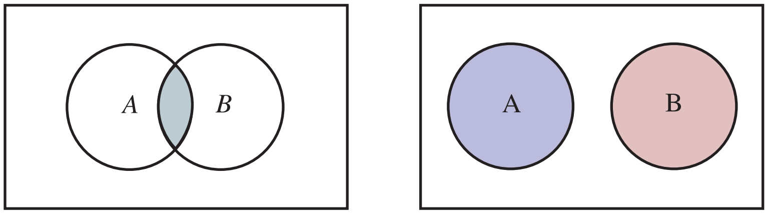
Saying that two events are mutually exclusive means that they can't both occur when the chance experiment is performed once.

It is sometimes useful to draw an informal picture of events to visualize relationships. In a **Venn diagram**, the collection of all possible outcomes is typically represented as the interior of a rectangle. Other events are then identified by specified regions inside this rectangle. Figure 6.3 illustrates several Venn diagrams.

Figure 6.3
Venn diagrams.



Not A　　*A or B*



A and B　　Mutually exclusive events

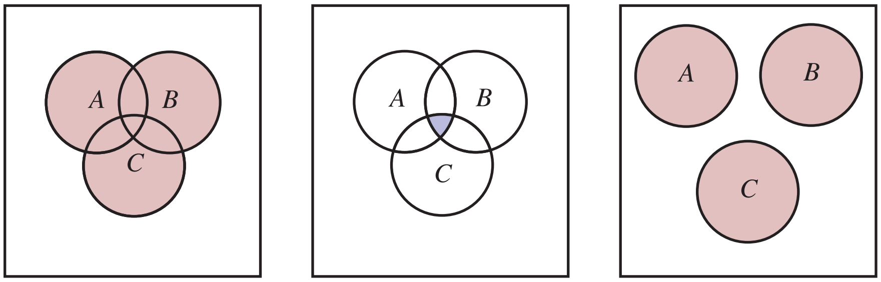
The use of the *or* and *and* operations can be extended to form new events from more than two events, as described in the following box and illustrated in Figure 6.4.

Let $A_1, A_2, \ldots, A_k$ denote k events.

1. The event A_1 *or* A_2 *or* … *or* A_k consists of all outcomes in at least one of the individual events $A_1, A_2, \ldots, A_k$.
2. The event A_1 *and* A_2 *and* … *and* A_k consists of all outcomes that are in every one of the individual events $A_1, A_2, \ldots, A_k$.

k events are mutually exclusive if no two of them have any common outcomes.

Figure 6.4
Venn diagrams with three events.



A or B or C　　*A and B and C*　　Three mutually exclusive events

Example 6.5 Asking Questions

A seminar class has four students. The instructor has an unusual way of asking questions. Four slips of paper numbered 1, 2, 3, and 4 are placed in a box. The instructor determines the student to whom any particular question will be addressed by selecting one of these four slips at random. Suppose that one question is to be posed during each of the next two class meetings. One possible outcome could be represented as (3, 1)—the first question is addressed to Student 3 and the second question to Student 1. There are 15 other possibilities. Consider the following events:

Copyright 2025 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s). Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

the event that the same student is asked both questions →

the event that Student 1 is asked at least one of the two questions →

$A = \{(1,1), (2,2), (3,3), (4,4)\}$
$B = \{(1,1), (1,2), (1,3), (1,4), (2,1), (3,1), (4,1)\}$
$C = \{(3,1), (2,2), (1,3)\}$
$D = \{(3,3), (3,4), (4,3)\}$
$E = \{(1,1), (1,3), (2,2), (3,1), (4,2), (3,3), (2,4), (4,4)\}$
$F = \{(1,1), (1,2), (2,1)\}$

Then

$$A \text{ or } C \text{ or } D = \{(1,1), (2,2), (3,3), (4,4), (3,1), (1,3), (3,4), (4,3)\}$$

The outcome (3,1) is contained in each of the events B, C, and E, as is the outcome (1,3). These are the only two common outcomes in all three events, so

$$B \text{ and } C \text{ and } E = \{(3,1), (1,3)\}$$

The events C, D, and F are mutually exclusive because no outcome in any one of these events is contained in either of the other two events.

Exercises 6.1–6.16

6.1 Define the term *chance experiment*, and give an example of a chance experiment with four possible outcomes.

6.2 Define the term *sample space*, and then give the sample space for the chance experiment you described in Exercise 6.1.

6.3 Consider the chance experiment in which the type of transmission—automatic (A) or manual (M)—is recorded for each of the next two cars purchased from a certain dealer.
- **a.** What is the set of all possible outcomes (the sample space)? (Hint: See Example 6.2.)
- **b.** Display the possible outcomes using a tree diagram.

6.4 Refer to the chance experiment described in the previous exercise and the sample space for that experiment.
- **a.** List the outcomes in each of the following events. Which of these events are simple events? (Hint: See Example 6.3.)
  - **i.** the event that at least one car has an automatic transmission
  - **ii.** the event that exactly one car has an automatic transmission
  - **iii.** the event that neither car has an automatic transmission
- **b.** What outcomes are in the event B *and* C? In the event B *or* C?

6.5 A tennis shop sells five different brands of rackets (Head, Prince, Slazenger, Wimbledon, and Wilson), each of which comes in either a midsize version or an oversize version. Consider the chance experiment in which brand and size are noted for the next racket purchased. One possible outcome is Head midsize, and another is Prince oversize.
- **a.** Let A denote the event that an oversize racket is purchased. List the outcomes in A.
- **b.** Let B denote the event that the name of the brand purchased begins with a W. List the outcomes in B.
- **c.** List the outcomes in the event *not* B.

6.6 Refer to the chance experiment described in the previous exercise.
- **a.** Head, Prince, and Wilson are U.S. companies and the others are not. Let C denote the event that the racket purchased is made by a U.S. company. List the outcomes in the event B *or* C. (Hint: See Example 6.4.)
- **b.** List outcomes in B *and* C.
- **c.** Display the possible outcomes using a tree diagram, with a first-generation branch for each brand.

6.7 A new model of laptop computer can be ordered with one of three screen sizes (10 inches, 12 inches, 15 inches) and one of three choices for the amount of RAM (8 GB, 16 GB, or 32 GB). Consider the chance experiment in which a laptop order is selected and the screen size and RAM choice are recorded.
- **a.** Display possible outcomes using a tree diagram.
- **b.** Let A be the event that the order is for a laptop with a screen size of 12 inches or smaller. Let B be the event that the order is for a laptop with at least 16 GB of RAM. What outcomes are in **i.** A^C? **ii.** $A \cup B$? **iii.** $A \cap B$?

Copyright 2025 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s). Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

c. Let C denote the event that the order is for a laptop with a 32 GB hard drive. Are A and C mutually exclusive events? Are B and C mutually exclusive?

6.8 At a particular college, students have three options for obtaining textbooks: purchase all textbooks, rent all textbooks for the semester, or purchase some books and rent some books. Consider a chance experiment in which a student from this college is selected at random and textbook option and class standing (freshman, sophomore, junior senior) are noted.

a. Give an example of a possible outcome for this chance experiment.

b. How many different outcomes would be in the sample space for this chance experiment?

c. Let A denote the event that the selected student rents all textbooks. What outcomes make up event A?

6.9 A bag of candy contains 2 pieces of milk chocolate and 3 pieces of dark chocolate. Suppose that these pieces are labeled M1, M2, D1, D2, D3. You plan to select pieces from the bag at random, stopping only when a dark chocolate piece has been selected.

a. Display the possible outcomes using a tree diagram.

b. What outcomes are contained in the event A, that exactly one piece is selected before the chance experiment terminates?

c. What outcomes are contained in the event C, that the chance experiment terminates with the third selection from the bag?

6.10 Suppose that, starting at a certain time, batteries coming off an assembly line are examined one by one to see whether they are defective (let D = defective and N = not defective). The chance experiment terminates as soon as a nondefective battery is obtained.

a. Give five possible outcomes for this chance experiment.

b. What can be said about the number of outcomes in the sample space?

c. What outcomes are in the event E, that the number of batteries examined is an even number?

6.11 Refer to the previous exercise and now suppose that the chance experiment terminates only when two nondefective batteries have been obtained.

a. Let A denote the event that at most three batteries must be examined before the chance experiment terminates. What outcomes are contained in A?

b. Let B be the event that exactly four batteries must be examined before the chance experiment terminates. What outcomes are in B?

c. What can be said about the number of possible outcomes for this chance experiment?

6.12 A family consisting of three people—P_1, P_2, and P_3—belongs to a medical clinic that always has a physician at each of stations 1, 2, and 3. During a certain week, each member of the family visits the clinic exactly once and is randomly assigned to a station. One outcome is (1, 2, 1), which means that P_1 is assigned to station 1, P_2 to station 2, and P_3 to station 1.

a. List the 27 possible outcomes. (Hint: First list the nine outcomes in which P_1 goes to station 1, then the nine in which P_1 goes to station 2, and finally the nine in which P_1 goes to station 3; a tree diagram might help.)

b. List all outcomes in the event A, that all three people go to the same station.

c. List all outcomes in the event B, that all three people go to different stations.

d. List all outcomes in the event C, that no one goes to station 2.

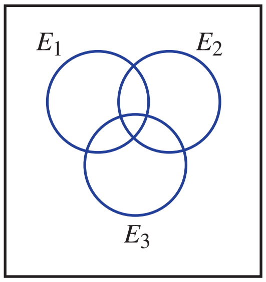
6.13 Using the outcomes for the chance experiment described in the previous exercise, identify outcomes in each of the following events:

a. B^C **b.** C^C **c.** $A \cup B$ **d.** $A \cap B$ **e.** $A \cap C$

6.14 An engineering construction firm is currently working on power plants at three different sites. Define events E_1, E_2, and E_3 as follows:

E_1 = the plant at Site 1 is completed by the contract date
E_2 = the plant at Site 2 is completed by the contract date
E_3 = the plant at Site 3 is completed by the contract date

The following Venn diagram can be used to picture the relationships among these events:



Shade the region in the Venn diagram corresponding to each of the following events. Redraw the Venn diagram for each part of the problem. (Hint: See the discussion of Venn diagrams.)

a. At least one plant is completed by the contract date.

b. All plants are completed by the contract date.

c. None of the plants are completed by the contract date.

6.15 For the events described in the previous exercise, shade the region in the Venn diagram corresponding to each of the following events. Redraw the Venn diagram for each part.

a. Only the plant at Site 1 is completed by the contract date.

b. Exactly one of the three plants is completed by the contract date.

c. Either the plant at Site 1 or both of the other two plants are completed by the contract date.

6.16 Consider a Venn diagram picturing two events A and B that are not mutually exclusive.

a. Shade the event $(A \cup B)^C$. On a separate Venn diagram shade the event $A^C \cap B^C$. How are these two events related?

b. Shade the event $(A \cap B)^C$. On a separate Venn diagram shade the event $A^C \cup B^C$. How are these two events related? (Note: These two relationships together are called DeMorgan's laws.)

Section 6.2 Definition of Probability

Reasoning about games of chance has a long history. Archeologists have found evidence that Egyptians used a small bone in mammals, the astragalus, as a sort of four-sided die as early as 3500 B.C. Games of chance were common in Greek and Roman times and during the Renaissance of Western Europe. From this early work related to games of chance, new interpretations of probability have evolved, including an empirical approach preferred by many statisticians today. The discussion of probability begins with a look at some of the different approaches to probability: the classical, relative frequency, and subjective approaches.

Classical Approach to Probability

Early mathematicians' development of the theory of probability was primarily in the context of gambling and games of chance. A characteristic common to most games of chance is a physical device used to produce outcomes. For example, in children's games, dice or a "spinner" might be used to create chance outcomes.

Dice are constructed so that the physical characteristics of each face are alike except for the number of dots. This ensures that the different outcomes for the roll of an individual die are equally likely. If this is the case, the probability of getting a five when a single six-sided die is rolled is one-sixth (1 chance in 6). In general, if there are N *equally likely* outcomes in the sample space for a chance experiment (such as rolling a die), the probability of each of the outcomes is $1/N$.

Classical Approach to Probability for Equally Likely Outcomes

When the outcomes in the sample space of a chance experiment are equally likely, the **probability of an event *E***, denoted by $\boldsymbol{P(E)}$, is the ratio of the number of outcomes favorable to E to the total number of outcomes in the sample space:

$$P(E) = \frac{\text{Number of outcomes favorable to } E}{\text{Number of outcomes in the sample space}}$$

According to this definition, the calculation of a probability consists of counting the number of outcomes that make up an event, counting the number of outcomes in the sample space, and then dividing.

Chance experiments that involve tossing fair coins, rolling fair dice, or selecting cards from a well-mixed deck of playing cards have equally likely outcomes. For example, if a fair die is rolled once, each outcome (simple event) has probability 1/6. With E denoting the event that the number rolled is even, $P(E) = 3/6$. This is just the number of outcomes in E (2, 4, and 6) divided by the total number of possible outcomes.

Example 6.6 Calling the Toss

On some football teams, the honor of calling the toss at the beginning of a football game is determined by random selection. Suppose that this week a member of the offense will call the toss. There are 5 linemen on the 11-player offense. If L represents the event that a lineman is selected to call the toss, 5 of the 11 possible outcomes are included in L. The probability that a lineman will be selected is then

$$P(L) = \frac{5}{11}$$

Example 6.7 Math Contest

Four students (Amit, Betina, Carlos, and Dara) submitted correct solutions to a math contest that had two prizes. Amit and Betina are math majors and Carlos and Dara are statistics majors. The contest rules specify that if more than two correct responses are submitted, the winners will be selected at random from those submitting correct responses. You can use (A, B) to denote the outcome that Amit and Betina are the two selected. The other possible outcomes can be denoted in a similar way. Then the sample space for the chance experiment of selecting the two winners from the four correct responses is

$$\{(A, B), (A, C), (A, D), (B, C), (B, D), (C, D)\}$$

Because the winners are selected at random, the six outcomes are equally likely and the probability of each individual outcome is 1/6.

Suppose that E denotes the event that both selected winners are the same major. Then

$$E = \{(A, B), (C, D)\}$$

Because E contains two outcomes, $P(E) = 2/6 = 0.333$.

If F denotes the event that at least one of the selected winners is a statistics major, then F consists of all outcomes except (A, B) and $P(F) = 5/6 = 0.833$.

What Do You Think?

Why might it be misleading to say that if a student is selected at random from the population of students at a four-year college, the probability that the selected student is a freshman is ¼?

The classical approach to probability works well for chance experiments that have a finite number of outcomes that are equally likely. However, many chance experiments do not have equally likely outcomes. For example, consider the chance experiment of selecting a student from those enrolled at a school and observing whether the student is a freshman, sophomore, junior, or senior. If there are more seniors than sophomores at the school, then the four possible outcomes for this chance experiment are not equally likely. In this situation, it would be a mistake to think that the probability of each outcome is ¼. To calculate or estimate probabilities in situations where outcomes are not equally likely, a different approach is needed.

Relative Frequency Approach to Probability

Early scientific investigators were aware that chance experiments do not always give the same results when repeated. However, even though it is not possible to predict the outcome of a chance experiment in any particular instance, there is a dependable and stable regularity when a chance experiment is repeated many times. This became the basis for fair wagering in games of chance.

For example, suppose that two friends, Chris and Tom, meet to play a game. A fair coin is flipped. If it lands heads up, Chris pays Tom $1; otherwise, Tom pays the same amount to Chris. After many repetitions, the proportion of the time that Chris wins will be close to one-half, and the two friends will have simply enjoyed the pleasure of each other's company for a few hours. This happy circumstance occurs because *in the long run* (meaning over many repetitions) the proportion of heads is 0.5. In the long run, half the time Chris wins, and half the time Tom wins.

When any given chance experiment is performed, some events are relatively likely to occur, whereas others are not as likely to occur. For a specified event E, a probability is a value that gives information about how likely it is that E will occur. In the relative frequency approach to probability, the probability describes how frequently E occurs when the chance experiment is performed many times.

Example 6.8 Tossing a Coin

Frequently, a coin is described as fair, or it is said that there is a 50% chance of a coin landing heads up. What is the meaning of the expression *fair*? It cannot refer to the result of a single toss, because a single toss cannot result in both a head and a tail. Might "fairness" and "50%" refer to 10 successive tosses yielding exactly five heads and five tails? Not really, because it is easy to imagine a fair coin landing heads up on only 3 or 4 of the 10 tosses.

Consider the chance experiment of tossing a coin just once. Define the simple events

H = event that the coin lands with its heads side facing up
T = event that the coin lands with its tails side facing up

Now suppose that you take a fair coin and begin to toss it over and over. After each toss, you calculate the relative frequency of heads observed so far. This calculation gives the value of the ratio

$$\frac{\text{number of times event } H \text{ occurs}}{\text{number of tosses}}$$

The results of the first 10 tosses might be as follows:

Toss	1	2	3	4	5	6	7	8	9	10
Cumulative number of H's	0	1	2	3	3	3	4	5	5	5
Relative frequency of H's	0	0.5	0.667	0.75	0.6	0.5	0.571	0.625	0.556	0.5

Figure 6.5 illustrates how the relative frequency of heads fluctuates during one sequence of 50 tosses. As the number of tosses increases, the relative frequency of heads does not continue to fluctuate wildly but instead stabilizes and approaches some fixed number (the limiting value). This stabilization is illustrated for a sequence of 1000 tosses in Figure 6.6.

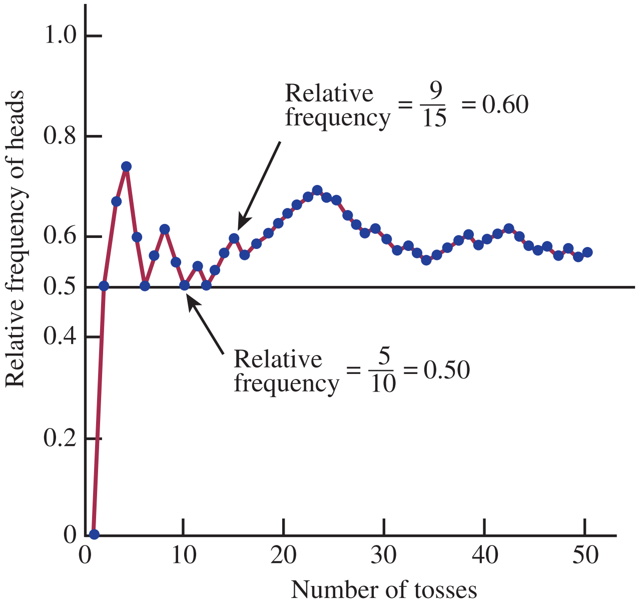
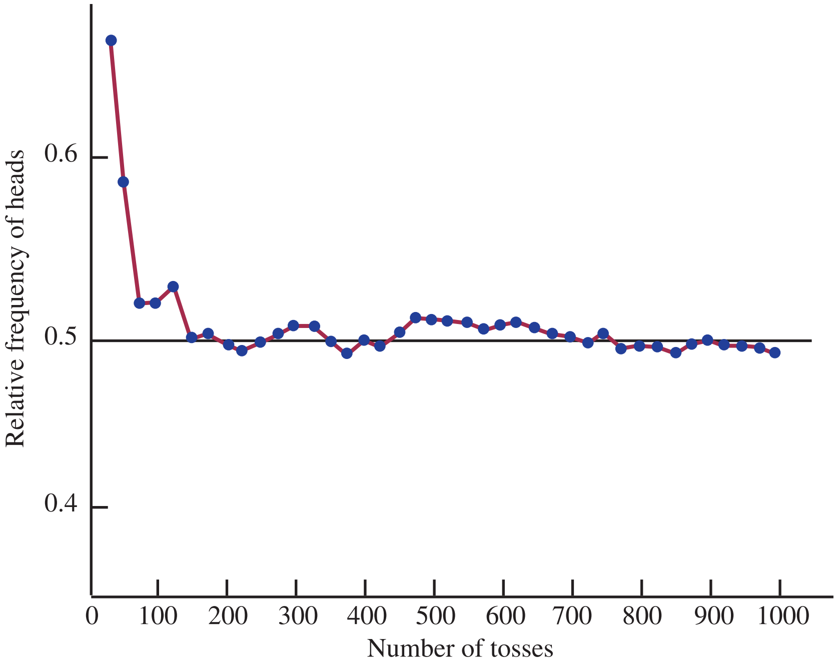


Figure 6.5
Relative frequency of heads in the first 50 of a long series of coin tosses.

Figure 6.6
Stabilization of the relative frequency of heads in coin tossing.



It may seem natural to you that the proportion of *H*'s would get closer and closer to the "real" probability of 0.5. That an empirically observed proportion could behave like this in real life also seemed reasonable to James Bernoulli in the early 18th century. He was able to prove mathematically what is now known as the law of large numbers.*

Law of Large Numbers

As the number of repetitions of a chance experiment increases, the chance that the relative frequency of occurrence for an event will differ from the true probability of the event by more than any small number approaches 0.

The law of large numbers tells you that, as the number of repetitions of a chance experiment increases, the proportion of the time an event *E* occurs gets close to and stays close to the real probability of *E* occurring even if the value of this probability is not known. This means that you can observe the outcomes of repetitions of a chance experiment and then use the observed outcomes to estimate probabilities.

Relative Frequency Approach to Probability

The **probability of an event *E***, denoted by ***P*(*E*)**, is defined to be the value approached by the relative frequency of occurrence of *E* when a chance experiment is performed many times. If the number of times the chance experiment is performed is quite large,

$$P(E) \approx \frac{\text{Number of times } E \text{ occurs}}{\text{Number of times the experiment is performed}}$$

The relative frequency definition of probability depends on being able to repeat a chance experiment under identical conditions. Suppose that you perform a chance experiment that consists of flipping a cap from a 20-ounce bottle of soda and noting whether the bottle cap lands with the open side up or down. The crucial difference from the previous coin-tossing

---

*Technically, this should be referred to more specifically as the weak law of large numbers for Bernoulli trials. After Bernoulli's proof, mathematical statisticians proved more general laws of large numbers.

Copyright 2025 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s). Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

chance experiment is that there is no particular reason to believe the cap is equally likely to land top up or top down.

If you assume that the chance experiment can be repeated under similar conditions (which seems reasonable), then you can flip the cap many times and calculate the relative frequency of the event *top up* (the proportion of the time the event has occurred so far):

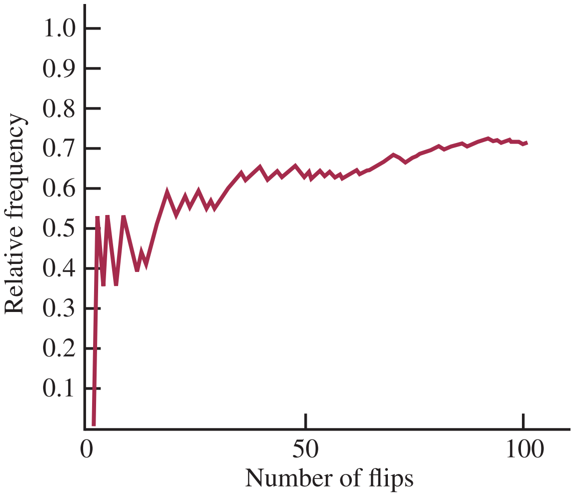
$$\frac{\text{number of times the event } \textit{top up} \text{ occurs}}{\text{number of flips}}$$

The results of the first 10 flips, with U indicating top up and D indicating top down, might be:

Flip	1	2	3	4	5	6	7	8	9	10
Outcome	U	U	D	U	D	D	U	D	U	U
Cumulative number of *ups*	1	2	2	3	3	3	4	4	5	6
Relative frequency of *ups*	1.0	1.0	0.67	0.75	0.6	0.5	0.57	0.5	0.56	0.6

Figure 6.7 illustrates how the relative frequency of the event *top up* fluctuates during a sequence of 100 flips. Based on these results and faith in the law of large numbers, it is reasonable to think that the probability of the cap landing top up is about 0.7.

Figure 6.7
Stabilization of the relative frequency of a bottle cap landing top up.



The relative frequency approach to probability is based on observation. From repeated observation, you can obtain stable relative frequencies that will, in the long run, provide reasonable estimates of the probabilities of different events. The relative frequency interpretation of probability is intuitive and widely used.

Subjective Approach to Probability

A third approach to probability is based on subjective judgments. In this view, a probability is interpreted as a personal measure of the strength of belief that a particular event will occur. A probability of 1 represents a belief that the event will certainly occur. A probability of 0 represents a belief that the event will certainly not occur—that it is impossible. Other probabilities are placed somewhere between 0 and 1, based on the strength of one's beliefs.

For example, an airline passenger might report their subjective assessment of the probability of being denied a seat as a result of overbooking on a particular flight as 0.01. Because this probability is close to 0, they believe it is very unlikely that they will be denied a seat.

The subjective interpretation of probability presents some difficulties. Different people may assign different probabilities to the same outcome, because each could have a different subjective belief. Although subjective probabilities are useful in studying and analyzing decision making, they are of limited use because they are personal and not generally replicable by others.

How Are Probabilities Determined?

Probability statements are frequently encountered in newspapers and magazines (and, of course, in this textbook). What is the basis for these probability statements? In most cases, research allows reasonable estimates based on observation and analysis. In general, if a probability is stated, it is based on one of the following approaches:

1. **The classical approach:** This approach is appropriate only for modeling chance experiments with equally likely outcomes.
2. **The subjective approach:** In this case, the probability represents an individual's judgment based on facts combined with personal evaluation of other information.
3. **The relative frequency approach:** A probability is estimated based on observing the outcomes of a chance experiment.

When you see probabilities in published sources, you should consider how the probabilities were produced. Be particularly cautious in the case of subjective probabilities.

Section 6.3 Basic Properties of Probability

Estimating probabilities by observing outcomes and then calculating a relative frequency is not always practical. The most obvious problem is time. If calculating probabilities even for simple chance experiments were to require hundreds or thousands of observations, everyone would avoid this task! An alternative approach that simplifies things in some situations is to study fundamental properties of probability that can be used to find probabilities of complex events. These fundamental properties are given in the accompanying box. A discussion of each property follows.

Fundamental Properties of Probability

1. For any event E, $0 \leq P(E) \leq 1$.
2. If S is the sample space for a chance experiment, $P(S) = 1$.
3. If two events E and F are mutually exclusive, then $P(E \text{ or } F) = P(E) + P(F)$.
4. For any event E, $P(E) + P(\text{not } E) = 1$. It follows that

$$P(\text{not } E) = 1 - P(E)$$

and

$$P(E) = 1 - P(\text{not } E).$$

Property 1: For Any Event E, $0 \leq P(E) \leq 1$

To understand the first property of probability, recall the previous bottle cap chance experiment where the number of flips landing with the top of the bottle cap facing up was observed. Suppose that you have flipped the cap N times. The number of times that the cap landed with the top facing up must be between 0 (it never happened) and N (it always happened).

Then the relative frequency of top up must be between two numbers:

$$\frac{0}{N} \leq \text{relative frequency} \leq \frac{N}{N}$$

This means that the relative frequency must always be greater than or equal to 0 and less than or equal to 1. As N increases, the long-run value of the relative frequency, which is the probability, must also lie between 0 and 1.

Copyright 2025 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s). Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

Property 2: If S Is the Sample Space for a Chance Experiment, $P(S) = 1$

The probability of any event is the proportion of time an outcome in the event will occur in the long run. Because the sample space consists of all possible outcomes for a chance experiment, in the long run an outcome that is in S will occur 100% of the time. This mean that $P(S) = 1$.

Property 3: If Two Events E and F Are Mutually Exclusive, then $P(E \text{ or } F) = P(E) + P(F)$

This is one of the most important properties of probability because it provides a method for calculating probabilities if the number of possible outcomes (simple events) in the sample space is finite. Any event is just a collection of outcomes from the sample space—some outcomes are in the event and others are not. Because simple events are mutually exclusive, any event can be viewed as a union of mutually exclusive events.

For example, suppose that a chance experiment has a sample space that consists of six outcomes, $O_1, O_2, \ldots, O_6$. Then $P(O_1)$ can be interpreted as the long run proportion of the time that the outcome O_1 will occur. The probabilities of the other outcomes can be interpreted in a similar way. Suppose that an event E is made up of outcomes O_1, O_2, and O_4. Then E will occur whenever O_1, O_2, or O_4 occurs. Because O_1, O_2, and O_4 are mutually exclusive, the long-run proportion of the time that E will occur is just the sum of the proportion of the time that each of the outcomes O_1, O_2, and O_4 will occur.

Because it is always possible to express any event as a collection of mutually exclusive simple events in this way, you can find the probability of an event by finding the probabilities of the simple events (outcomes) that make up the event and then adding. If the probabilities of all the simple events are known, it is then easy to calculate the probability of more complex events.

Property 4: For Any Event E, $P(E) + P(\textit{not } E) = 1$. It Follows that $P(\textit{not } E) = 1 - P(E)$ and $P(E) = 1 - P(\textit{not } E)$

Property 4 follows from Properties 2 and 3. Property 3 says that $P(E)$ is the sum of the probabilities of the simple events that are in E and that $P(\textit{not } E)$ is the sum of the probabilities for simple events that are in *not* E. Every simple event in the sample space is in either E or *not* E. Property 2 says that the sum of the probabilities of all the simple events is 1. It follows that $P(E) + P(\textit{not } E)$ must equal 1.

Property 4 implies that $P(E) = 1 - P(\textit{not } E)$. Knowing this is surprisingly useful. There are many situations where it is easier to calculate $P(\textit{not } E)$ than it is to calculate $P(E)$ directly.

Example 6.9 Preventable Deaths

Understand the context ➤

The authors of the paper **"Preventable Deaths Due to Problems in Care in English Acute Hospitals" (*BMJ Quality & Safety* [2012]: 737–745)** carried out a study of hospital deaths. They estimated that 5.2% of the deaths that occurred in English hospitals were preventable. They attributed the preventable deaths to the following types of errors: (1) poor patient monitoring; (2) diagnostic errors; (3) inadequate drug or fluid management; and (4) other. It was estimated that for the preventable deaths

31% were due to poor patient monitoring
30% were due to diagnostic errors
21% were due to inadequate drug or fluid management
18% were due to other errors

Consider the chance experiment that consists of observing the type of error for a randomly selected preventable death. The simple events for this experiment can be represented by O_1, O_2, O_3, and O_4:

O_1 = the simple event that the death was due to poor patient monitoring
O_2 = the simple event that the death was due to a diagnostic error

O_3 = the simple event that the death was due to inadequate drug or fluid management
O_4 = the simple event that the death was due to other errors

The following table displays the probabilities for the simple events based on the estimates given in the paper.

Simple Event (Outcome)	O_1	O_2	O_3	O_4
Probability	0.31	0.30	0.21	0.18

Consider the event E, the event that *a randomly selected preventable death is due to poor patient monitoring, a diagnostic error, or inadequate drug or fluid control.* This event consists of the outcomes O_1, O_2, and O_3. It follows that

$$\begin{aligned} P(E) &= P(O_1 \cup O_2 \cup O_3) \\ &= P(O_1) + P(O_2) + P(O_3) \\ &= 0.31 + 0.30 + 0.21 \\ &= 0.82 \end{aligned}$$

That is, in the long run, 82% of all preventable deaths are due to poor patient monitoring, diagnostic errors, or inadequate drug or fluid monitoring. Another way to calculate this probability is to notice that the event *not E* consists only of the outcome O_4. Another way to calculate $P(E)$ is

$$P(E) = 1 - P(\textit{not } E) = 1 - P(O_4) = 1 - 0.18 = 0.82$$

Having established the fundamental properties of probability, you are now ready to consider some probability rules that can be used to evaluate probabilities in some situations. An important aspect of each rule is the set of conditions necessary in order to use the rule. A common error is to believe that the rules can be used in *any* probability calculation. This is *not* the case. You must be careful to use the rules only after verifying that any necessary conditions are met.

Equally Likely Outcomes

The first probability rule considered applies only when the simple events in the sample space are equally likely. Chance experiments involving tossing fair coins, rolling fair dice, or selecting cards from a well-mixed deck of playing cards are examples of chance experiments that have equally likely outcomes. If a fair die is rolled once, each possible outcome (1, 2, 3, 4, 5, and 6) has probability 1/6. With E denoting the event that the outcome is an even number,

$$P(E) = P(2 \textit{ or } 4 \textit{ or } 6) = P(2) + P(4) + P(6) = \frac{1}{6} + \frac{1}{6} + \frac{1}{6} = \frac{3}{6}$$

This is just the ratio of the number of outcomes in E to the total number of possible outcomes. The following box presents the generalization of this result.

Calculating Probabilities When Simple Events in the Sample Space Are Equally Likely

Consider a chance experiment that can result in any one of N possible outcomes. Denote the corresponding simple events by $O_1, O_2, \dots, O_N$. If these simple events are equally likely to occur, then

1. $P(O_1) = \frac{1}{N}, P(O_2) = \frac{1}{N}, \dots, P(O_N) = \frac{1}{N}$
2. For *any* event E,

$$P(E) = \frac{\text{number of simple events in } E}{N}$$

Copyright 2025 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s). Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

Addition Rule for Mutually Exclusive Events

You have seen previously that the probability of an event can be calculated by adding together probabilities of the simple events that correspond to the outcomes making up the event. This addition process can also be used to calculate the probability of the union of two events that are mutually exclusive.

The Addition Rule for Mutually Exclusive Events

Suppose E and F are mutually exclusive events. Then

$$P(E \text{ or } F) = P(E \cup F) = P(E) + P(F)$$

This property of probability is known as the addition rule for mutually exclusive events. More generally, if events $E_1, E_2, \ldots, E_k$ are all mutually exclusive, then

$$P(E_1 \text{ or } E_2 \text{ or} \ldots \text{or } E_k) = P(E_1 \cup E_2 \cup \cdots \cup E_k) = P(E_1) + P(E_2) + \cdots + P(E_k)$$

In words, the probability that at least one of these k mutually exclusive events occurs is the sum of the probabilities of the individual events.

Kirilldz/Shutterstock.com

Consider the chance experiment that consists of rolling a pair of fair dice. Because a fair die has six sides, each of which is equally likely to land on the top side, there are 36 possible outcomes for this chance experiment, such as (1, 1), (1, 2), and so on. The dice are fair, so these outcomes are equally likely. You can calculate the probability of observing a total of 5 on the two dice by counting the number of outcomes that result in a total of 5. There are four such outcomes—(1, 4), (2, 3), (3, 2), and (4, 1). Then

$$P(\text{total of } 5) = \frac{4}{36}$$

The probabilities for the other possible totals can be calculated in a similar way, and they are shown in the following table:

Total	Probability	Total	Probability
2	1/36	8	5/36
3	2/36	9	4/36
4	3/36	10	3/36
5	4/36	11	2/36
6	5/36	12	1/36
7	6/36		

What is the probability of getting a total of 3 or 5? Consider the two events

$$E = \text{total is } 3$$

and $F = \text{total is } 5$

Clearly E and F are mutually exclusive because the sum cannot simultaneously be both 3 and 5. Notice also that neither of these events is simple, because neither consists of only a single outcome from the sample space. You can apply the addition rule for mutually exclusive events as follows:

$$P(E \text{ or } F) = P(E \cup F) = P(E) + P(F) = \frac{2}{36} + \frac{4}{36} = \frac{6}{36}$$

Example 6.10 Car Choices

Understand the context ➤

A large auto center sells cars made by a number of different manufacturers. Three of these manufacturers are Japanese: Honda, Nissan, and Toyota. Consider a chance experiment that consists of observing the make and model of the next car purchased at this auto center.

Copyright 2025 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s). Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

The outcomes in the sample space would then be simple events such as Nissan Altima and Toyota Prius. Let's define events E_1, E_2, and E_3 by

$$E_1 = \text{Honda}$$
$$E_2 = \text{Nissan}$$
$$\text{and} \quad E_3 = \text{Toyota}$$

Notice that E_1 is not a simple event because there is more than one model Honda (for example, Civic, Accord, and Fit).

Suppose that based on several years of sales data, the probabilities of these three events have been estimated empirically as $P(E_1) = 0.25$, $P(E_2) = 0.18$, and $P(E_3) = 0.14$. Because E_1, E_2, and E_3 are mutually exclusive, the addition rule gives

$$\begin{aligned} P(\text{Honda } \textit{or} \text{ Nissan } \textit{or} \text{ Toyota}) &= P(E_1 \cup E_2 \cup E_3) \\ &= P(E_1) + P(E_2) + P(E_3) \\ &= 0.25 + 0.18 + 0.14 = 0.57 \end{aligned}$$

The probability that the next car purchased is *not* made by one of these three manufacturers is

$$P(\textit{not}(E_1 \textit{ or } E_2 \textit{ or } E_3)) = P((E_1 \cup E_2 \cup E_3)^C) = 1 - 0.57 = 0.43$$

In Section 6.6, you will see how $P(E \cup F)$ can be calculated when the two events E and F are not mutually exclusive.

Exercises 6.17–6.37

6.17 A large department store with multiple locations offers online ordering. When a purchase is made online, the customer can select one of four different delivery options: expedited overnight delivery, expedited second-day delivery, standard delivery, or delivery to the nearest store for customer pick-up. Consider the chance experiment that consists of observing the selected delivery option for a randomly selected online purchase. What are the four simple events that make up the sample space for this experiment?

6.18 Consider the chance experiment described in the previous exercise. Suppose that the probability of an overnight delivery selection is 0.1, the probability of a second-day delivery selection is 0.3, and the probability of a standard-delivery selection is 0.4. Find the following probabilities. (Hint: See Example 6.9.)

a. the probability that a randomly selected online purchase selects delivery to the nearest store for customer pick-up.

b. the probability that the customer selects a form of expedited delivery.

c. the probability that either standard delivery or delivery to the nearest store is selected.

6.19 The article **"Scrambled Statistics: What Are the Chances of Finding Multi-Yolk Eggs?" (*Significance* [August 2016]: 11)** gives the probability of a double-yolk egg as 0.001.

a. Give a relative frequency interpretation of this probability.

b. If 5000 eggs were randomly selected, about how many double-yolk eggs would you expect to find?

6.20 For a monthly subscription fee, a video download site allows people to download and watch up to five movies per month. Based on past download histories, the following table gives the estimated probabilities that a randomly selected subscriber will download 0, 1, 2, 3, 4, or 5 movies in a particular month.

Number of Downloads	0	1	2	3	4	5
Estimated Probability	0.03	0.45	0.25	0.10	0.10	0.07

If a subscriber is selected at random, what is the estimated probability that this subscriber downloads

a. three or fewer movies?

b. at most three movies?

c. four or more movies?

d. zero or one movie?

e. more than one movie?

Copyright 2025 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s). Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

6.21 A bookstore sells two types of books (fiction and nonfiction) in several formats (hardcover, paperback, digital, and audio). For the chance experiment that consists of observing the type and format of a single-book purchase, two of the eight possible outcomes are a hardcover fiction book and an audio nonfiction book.

a. There are eight outcomes in the sample space for this experiment. List these possible outcomes.

b. Do you think it is reasonable to think that the outcomes for this experiment would be equally likely? Explain.

6.22 Consider the chance experiment described in the previous exercise.

a. For customers who purchase a single book, the estimated probabilities for the different possible outcomes are given in the cells of the accompanying table. What is the probability that a randomly selected single-book purchase will be for a book in print format (hardcover or paperback)?

	Hardcover	**Paperback**	**Digital**	**Audio**
Fiction	0.15	0.45	0.10	0.10
Nonfiction	0.08	0.04	0.02	0.06

b. Show two different ways to calculate the probability that a randomly selected single-book purchase will be for a book that is not in a print format. (Hint: See Example 6.9.)

c. Calculate the probability that a randomly selected single-book purchase will be for a work of fiction.

6.23 Medical insurance status—covered (C) or not covered (N)—is determined for each individual arriving for treatment at a hospital's emergency room. Consider the chance experiment in which this determination is made for two randomly selected patients.

The simple events for this chance experiment are $O_1 = (C, C)$, meaning that the first patient selected was covered and the second patient selected was also covered, $O_2 = (C, N)$, $O_3 = (N, C)$, and $O_4 = (N, N)$. Suppose that probabilities are $P(O_1) = 0.81$, $P(O_2) = 0.09$, $P(O_3) = 0.09$, and $P(O_4) = 0.01$.

a. Define A as the event that at most one patient is covered.

i. What simple events are in the event A?

ii. Calculate $P(A)$.

b. Define B as the event that the two patients have the same coverage status.

i. What simple events are in the event B?

ii. Calculate $P(B)$.

6.24 Roulette is a game of chance that involves spinning a wheel that is divided into 38 equal segments, as shown in the accompanying picture.

Mindscape studio/Shutterstock.com

A metal ball is tossed into the wheel as it is spinning, and the ball eventually lands in one of the 38 segments. Each segment has an associated color. Two segments are green. Half of the other 36 segments are red and the others are black. When a balanced roulette wheel is spun, the ball is equally likely to land in any one of the 38 segments.

a. When a balanced roulette wheel is spun, what is the probability that the ball lands in a red segment?

b. In the roulette wheel shown, black and red segments alternate. Suppose instead that the red segments were side-by-side and that the black segments were together. Does this increase the probability that the ball will land in a red segment? Explain.

c. Suppose that you watch 1000 spins of a roulette wheel and note the color that results from each spin. What would be an indication that the wheel was not balanced?

6.25 Phoenix is a hub for a large airline. Suppose that on a particular day, 8000 passengers arrived in Phoenix on this airline. Phoenix was the final destination for 1800 of these passengers. The others were all connecting to flights to other cities. On this particular day, several inbound flights were late, and 480 connecting passengers missed their connecting flight and were delayed in Phoenix. Of the 480 who were delayed, 75 were delayed overnight and had to spend the night in Phoenix. Consider the chance experiment of choosing a passenger at random from these 8000 passengers. Calculate the following probabilities:

a. the probability that the selected passenger had Phoenix as a final destination.

b. the probability that the selected passenger did not have Phoenix as a final destination.

c. the probability that the selected passenger was connecting and missed the connecting flight.

Copyright 2025 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s). Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

d. the probability that the selected passenger was a connecting passenger and did not miss the connecting flight.
e. the probability that the selected passenger either had Phoenix as a final destination or was delayed overnight in Phoenix.

6.26 A customer satisfaction survey is planned. The company carrying out the survey plans to contact 50 passengers selected at random from the 8000 passengers who arrived in Phoenix on the day described in the previous exercise. The airline knows that the survey results will not be favorable if too many people who were delayed overnight are included in the survey. Should the airline be worried? Write a few sentences explaining whether or not you think the airline should be worried, using relevant probabilities to support your answer.

6.27 A professor assigns five problems to be completed as homework. At the next class meeting, two of the five problems will be selected at random and collected for grading. You only completed the first three problems.
a. What is the probability that you will be able to turn in both of the problems selected? (Hint: You can think of the problems as being labeled A, B, C, D, and E. Then one possible selection of two problems is A and B. If these are the two problems selected and you did problems A, B, and C, you will be able to turn in both problems. There are nine other possible selections to consider.)
b. Does the probability that you will be able to turn in both problems change if you had completed the last three problems instead of the first three problems? Explain.
c. What happens to the probability that you will be able to turn in both problems selected if you had completed four of the problems rather than just three?

6.28 The **National Center for Health Statistics (cdc.gov)** gave the following information on births in the United States in 2019:

Type of Birth	Number of Births
Single child	3,623,963
Twins	120,291
Triplets	3,136
Quadruplets	114
Quintuplets or higher	36

Use this information to estimate the probability that a randomly selected birth in 2019 was
a. twins
b. quadruplets
c. more than a single child

6.29 The report **"Teens, Social Media & Technology Overview 2015" (Pew Research Center, April 9, 2015)** summarized data from a large survey of teens age 13 to 17. Of those surveyed, 71% use Facebook and 52% use Instagram. Use these percentages to explain why the two events identified below cannot be mutually exclusive events.

F = event that a randomly selected survey participant uses Facebook
and
I = event that a randomly selected survey participant uses Instagram

6.30 According to ***The Chronicle for Higher Education Almanac* (2021)**, in a 2019 survey of 95,505 first-time freshmen enrolled at four-year colleges, 52,623 reported that they were enrolled at their first-choice college.
a. If a person who participated in this survey is selected at random, what is the probability that the selected person was enrolled at their first-choice college?
b. What is the probability that the selected person was not enrolled at their first-choice college?

6.31 The same issue of ***The Chronicle for Higher Education Almanac*** referenced in the previous exercise also reported the following information for first-time freshmen at four-year colleges in 2019:

- A total of 95,505 first-time freshmen were surveyed.
- 52,623 were enrolled in their first-choice college.
- 26,357 were enrolled at their second choice college.
- The rest were enrolled at a college that was not their first- or second-choice college.

What is the probability that a randomly selected person who participated in this survey
a. was enrolled in their second-choice college?
b. was not enrolled at their first-choice college?
c. was not enrolled at either their first- or second-choice college?

6.32 A deck of 52 playing cards is mixed well, and 5 cards are dealt. A standard deck of cards includes 13 cards of each of four card types (suits): clubs, diamonds, hearts, and spades.
a. It can be shown that (disregarding the order in which the cards are dealt) there are 2,598,960 possible five-card hands, of which only 1287 are hands consisting entirely of spades. What is the probability that a hand will consist entirely of spades? What is the probability that a hand will consist entirely of a single suit?
b. It can be shown that exactly 63,206 hands contain only spades and clubs, with both suits represented. What is the probability that a hand consists entirely of spades and clubs with both suits represented?

Copyright 2025 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s). Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

c. Using the result of Part (b), what is the probability that a hand contains cards from exactly two suits?

6.33 Suppose you want to estimate the probability that a patient will develop an infection while hospitalized at a particular hospital. In the past year, this hospital had 6450 patients, and 712 of them developed an infection. What is the estimated probability that a patient at this hospital will develop an infection?

6.34 Six people hope to be selected as a contestant on a TV game show. Two of these people are younger than 25 years old. Two of these six will be chosen at random to be on the show.

a. What is the sample space for the chance experiment of selecting two of these people at random? (Hint: You can think of the people as being labeled A, B, C, D, E, and F. One possible selection of two people is A and B. There are 14 other possible selections to consider.)

b. Are the outcomes in the sample space equally likely?

c. What is the probability that both chosen contestants are younger than 25?

d. What is the probability that both chosen contestants are not younger than 25?

e. What is the probability that one is younger than 25 and the other is not?

6.35 The student council for a school of science and math has one representative from each of the five academic departments: biology (B), chemistry (C), mathematics (M), physics (P), and statistics (S). Two of these students are to be randomly selected for a university-wide student committee. The two students will be selected by placing five slips of paper (numbered 1, 2, 3, 4, and 5) into a bowl, mixing, and drawing out two of them.

a. What are the 10 possible outcomes (simple events)?

b. From the description of the selection process, all simple events are equally likely. What is the probability of each simple event?

c. What is the probability that one of the committee members is the statistics department representative?

d. What is the probability that both committee members come from laboratory science departments?

6.36 A student placement center has requests from five students for interviews for employment with a consulting firm. Three of these students are math majors, and the other two students are statistics majors. Unfortunately, the interviewer only has time to talk with two of the students. These two will be randomly selected from among the five.

a. What is the probability that both selected students are statistics majors?

b. What is the probability that both students selected are math majors?

c. What is the probability that at least one of the students selected is a statistics major?

d. What is the probability that the selected students have different majors?

6.37 Suppose that a six-sided die is weighted so that any even-numbered face is twice as likely to land face up as any odd-numbered face. Consider the chance experiment that consists of rolling this die.

a. What are the probabilities of the six simple events? (Hint: Denote these events by $O_1, \dots, O_6$. Then $P(O_1) = p$, $P(O_2) = 2p$, $P(O_3) = p, \dots,$ $P(O_6) = 2p$. Now use a condition on the sum of these probabilities to determine p.)

b. What is the probability that the number showing is an odd number? What is the probability that the number showing is at most three?

c. Now suppose that the die is weighted so that the probability of each simple event is proportional to the number showing on the corresponding upturned face; that is, $P(O_1) = c$, $P(O_2) = 2c, \dots,$ $P(O_6) = 6c$. What are the probabilities of the six simple events? Calculate the probabilities of Part (b) for this die.

Section 6.4 Conditional Probability

Sometimes the knowledge that one event has occurred changes the assessment of the likelihood that another event occurs. For example, consider a population in which 0.1% of all individuals have a certain disease. It is difficult to tell if someone has the disease based on a physical exam, but there is a diagnostic test available. Unfortunately, the test is not always correct. Of those with positive test results, 80% actually have the disease and the other 20% who show positive test results are false-positives.

To put this in probability terms, consider the chance experiment of randomly selecting a person from the population. Define the following events:

E = event that the individual has the disease
F = event that the individual's diagnostic test is positive

Copyright 2025 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s). Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

$P(E|F)$ is used to denote the probability of the event E *given that* the event F is known to have occurred. A new symbol has been used to indicate that a probability calculation has been made conditional on the occurrence of another event. The standard symbol for this is a vertical line, and it is read "given." For example, you would say, "the probability that an individual has the disease *given* that the diagnostic test is positive." This is represented symbolically as $P(\text{has disease}|\text{positive test})$ or $P(E|F)$. This probability is called a **conditional probability**.

The information provided about the disease and the diagnostic test implies that

$$P(E) = 0.001$$
$$P(E|F) = 0.8$$

This means that before you have diagnostic test information, the occurrence of E is unlikely. However, once it is known that the test result is positive, the likelihood of the disease increases dramatically. (If this were not so, the diagnostic test would not be very useful!)

Example 6.11 Racial Disparities in Searches During Traffic Stops

Understand the context ➤

Justice Navigator is an online tool that provides access to policing data for some cities. It can be used to investigate racial inequities in policing. Data from Justice Navigator for the city of Sacramento, California, on traffic stops made in 2019 are given in the table below. Each of the 36,716 traffic stops made by police in 2019 was classified according to the race of the person stopped and whether or not the stop resulted in a search of the driver or the vehicle.

Consider the data ➤

	Race of Driver					
	Asian	**Black**	**Latinx**	**White**	**Other**	**Total**
Searched	146	2,822	1,136	800	94	**4,998**
Not Searched	2,163	11,034	7,268	9,827	1,426	**31,718**
Total	**2,309**	**13,856**	**8,404**	**10,627**	**1,520**	**36,716**

Look to see how each of the following statements follows from the information given in the table:

1. There were 800 White drivers who were searched.
2. The total number of people who were not searched was 31,718.
3. There were 8,404 Latinx drivers stopped.
4. The total number of traffic stops was 36,716.

Do the work ➤

Suppose that one of the 2019 traffic stops is selected at random. Because the stop will be selected at random, each of the 36,716 stops is equally likely to be selected. What is the probability that a randomly selected stop involved a Latinx driver? You can calculate this probability as follows:

$$P(\text{driver is Latinx}) = \frac{\text{number of Latinx drivers stopped}}{\text{total number of drivers stopped}} = \frac{8{,}404}{36{,}716} = 0.23$$

This means that 23% of the traffic stops involved Latinx drivers.

You can also calculate the probabilities that a randomly selected stop involved a driver for each of the other groups for which data was reported:

$$P(\text{driver is Asian}) = \frac{2{,}309}{36{,}716} = 0.06$$

$$P(\text{driver is Black}) = \frac{13{,}856}{36{,}716} = 0.38$$

$$P(\text{driver is White}) = \frac{10{,}677}{36{,}716} = 0.29$$

What Do You Think?

In Sacramento in 2019, 14% of the population was Black, 20% of the population was Asian, 30% of the population was Latinx, and 34% of the population was White. How is this information relevant to assessing racial disparities in traffic stops?

These probabilities suggest a racial disparity in the likelihood of being pulled over in a traffic stop. Thirty-eight percent of drivers stopped were Black, which is much greater than the percentage of the population who are Black, which is 14%. The percentage of drivers stopped who were Latinx is also higher than corresponding population percentage. The percentages of drivers stopped who were Asian or White were lower than the corresponding population percentages.

Some traffic stops result in searches of the driver or the vehicle. The probability of a 2019 traffic stop in Sacramento resulting in a search is

$$P(search) = \frac{\text{number of drivers searched}}{\text{total number of stops}} = \frac{4{,}998}{36{,}716} = 0.14$$

This means that 14% of the traffic stops resulted in a search.

Now suppose you know that the driver stopped is Black. Does this information change the assessment of the likelihood that the stop will result in a search? Because 2,822 of the 13,856 Black drivers stopped were searched, this suggests that

$$P(search \mid Black) = \frac{\text{number of Black drivers searched}}{\text{total number of Black drivers stopped}} = \frac{2{,}822}{13{,}856} = 0.20$$

The probability is calculated this way because you know that the driver stopped is one of the 13,856 Black drivers who were stopped, each of whom is equally likely to have been selected.

Also consider the following conditional probabilities:

$$P(search \mid Asian) = \frac{146}{2309} = 0.06$$

$$P(search \mid Latinx) = \frac{1136}{8404} = 0.14$$

$$P(search \mid White) = \frac{800}{10{,}627} = 0.08$$

Interpret the results P ➤

Notice that the probability that a stop resulted in a search differs depending on the race of the driver. For example, if you were to repeat the chance experiment of selecting a driver at random, about 20% *of the times that resulted in a Black driver being selected* would also result in the selection of a driver who was searched. Notice that the unconditional probability $P(\text{Search}) = 0.14$ is not equal to the conditional probability $P(\text{Search}|\text{Black}) = 0.20$, again suggesting racial disparities in terms of which drivers are searched. This context will be revisited in Example 6.14, which looks at the outcomes resulting from the searches.

Example 6.12 GFI Switches

Understand the context ➤

A GFI (ground fault interrupt) switch turns off power to a system in the event of an electrical malfunction. A spa manufacturer currently has 25 spas in stock, each equipped with a single GFI switch. Suppose that two different companies supplied the switches, and that some of the switches are defective, as summarized in the following table:

Consider the data ➤

	Nondefective	Defective	Total
Company 1	10	5	**15**
Company 2	8	2	**10**
Total	**18**	**7**	**25**

A spa is randomly selected for testing. Consider the following events:

E = event that GFI switch in the selected spa is from Company 1
F = event that GFI switch in the selected spa is defective

Do the work ➤ Using the information in the table, you can calculate the following probabilities:

$$P(E) = \frac{15}{25} = 0.60 \quad P(F) = \frac{7}{25} = 0.28 \quad P(E \text{ and } F) = P(E \cap F) = \frac{5}{25} = 0.20$$

Now suppose that testing reveals a defective switch. This means that the chosen spa is one of the seven in the "defective" column. How likely is it that the switch came from the first company? Because five of the seven defective switches are from Company 1,

$$P(E|F) = P(\text{company } 1|\text{defective}) = \frac{5}{7} = 0.714$$

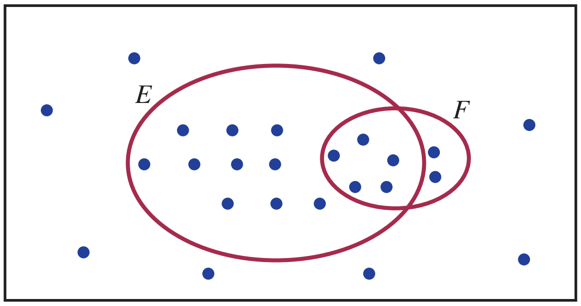
Notice that this is larger than the unconditional probability $P(E)$. This is because Company 1 has a much higher defective rate than Company 2.

An alternative expression for the conditional probability is

$$P(E|F) = \frac{5}{7} = \frac{5/25}{7/25} = \frac{P(E \text{ and } F)}{P(F)} = \frac{P(E \cap F)}{P(F)}$$

Notice that $P(E|F)$ is a ratio of two previously specified probabilities. It is the probability that both events occur divided by the probability of the "conditioning event" F. Additional insight comes from the Venn diagram of Figure 6.8. Once it is known that the outcome lies in F, the chance that E also occurs is the "size" of (E *and* F) relative to the size of F.

Figure 6.8
Venn diagram for Example 6.12 (each dot represents one GFI switch).



The previous example leads to a general definition of conditional probability.

Definition

Conditional probability: Suppose that E and F are two events with $P(F) > 0$. **The conditional probability of the event E given that the event F has occurred**, denoted by $P(E|F)$, is

$$P(E|F) = \frac{P(E \cap F)}{P(F)}$$

Notice the requirement that $P(F) > 0$. In addition to the standard warning about division by 0, there is another reason for requiring $P(F)$ to be positive. If the probability of F were 0, the event F would never occur. Therefore, it would not make sense to calculate the probability of another event conditional on F having occurred.

Example 6.13 Surviving a Heart Attack

Understand the context ➤ Medical guidelines recommend that a hospitalized patient who suffers a cardiac arrest (a serious condition where the heart stops beating) should receive defibrillation (an electric

Copyright 2025 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s). Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

shock to the heart) within 2 minutes. The paper **"Delayed Time to Defibrillation After In-Hospital Cardiac Arrest" (*The New England Journal of Medicine* [2008]: 9–17)** describes a study of the time to defibrillation for hospitalized patients in hospitals of various sizes.

The authors examined medical records of 6716 patients who suffered cardiac arrest while hospitalized, recording the size of the hospital and whether or not defibrillation occurred in 2 minutes or less. Data from this study are summarized in the accompanying table.

Consider the data ➤

	Time to Defibrillation		
Hospital Size	**2 minutes or less**	**More than 2 minutes**	**Total**
Small (Less than 250 beds)	1124	576	**1700**
Medium (250–499 beds)	2178	886	**3064**
Large (500 or more beds)	1387	565	**1952**
Total	**4689**	**2027**	**6716**

Suppose that these data are representative of the larger group of all hospitalized patients who suffer a cardiac arrest and that a hospitalized patient who suffered a cardiac arrest is selected at random. The following events are of interest:

S = event that the selected patient is at a small hospital
M = event that the selected patient is at a medium-sized hospital
L = event that the selected patient is at a large hospital
D = event that the selected patient receives defibrillation in 2 minutes or less

Do the work ➤ You can use the information in the table to calculate

$$P(D) = \frac{4689}{6716} = 0.698$$

This probability is interpreted as the proportion of hospitalized patients who suffer cardiac arrest that receive defibrillation in 2 minutes or less. This means that 69.8% of these patients receive timely defibrillation.

Now suppose it is known that the selected patient was at a small hospital. How likely is it that this patient received defibrillation in 2 minutes or less? To answer this question, you need to calculate $P(D|S)$, the probability of defibrillation in 2 minutes or less *given* that the patient is at a small hospital. Using the formula for conditional probability, you know

$$P(D|S) = \frac{P(D \cap S)}{P(S)}$$

From the information in the table, you can calculate

$$P(D \cap S) = \frac{1124}{6716} = 0.167$$

$$P(S) = \frac{1700}{6716} = 0.253$$

and so

$$P(D|S) = \frac{P(D \cap S)}{P(S)} = \frac{0.167}{0.253} = 0.660$$

Notice that this is smaller than the unconditional probability, $P(D) = 0.698$. This tells you that there is a smaller probability of timely defibrillation at a small hospital.

Two other conditional probabilities of interest are

$$P(D|M) = \frac{P(D \cap M)}{P(M)} = \frac{2178/6716}{3064/6716} = 0.711$$

and

$$P(D|L) = \frac{P(D \cap L)}{P(L)} = \frac{1387/6716}{1952/6716} = 0.711$$

From this, you can see that the probability of timely defibrillation is the same for patients at medium-sized and large hospitals, and that this probability is higher than that for patients at small hospitals.

It is also possible to calculate $P(L|D)$, the probability that a patient is at a large hospital given that the patient received timely defibrillation:

$$P(L|D) = \frac{P(L \cap D)}{P(D)} = \frac{1387/6716}{4689/6716} = 0.296$$

Interpret the results ➤ Look carefully at the interpretation of some of these probabilities:

1. $P(D) = 0.698$ is interpreted as the proportion *of all hospitalized patients* suffering cardiac arrest who receive timely defibrillation. Approximately 69.8% of these patients receive timely defibrillation.
2. $P(D \cap L) = \frac{1387}{6716} = 0.207$ is the proportion *of all hospitalized patients* who suffer cardiac arrest who are at a large hospital *and* who receive timely defibrillation.
3. $P(D|L) = 0.711$ is the proportion *of patients at large hospitals* suffering cardiac arrest who receive timely defibrillation.
4. $P(L|D) = 0.296$ is the proportion *of patients who receive timely defibrillation* who were at large hospitals.

Notice the difference between the unconditional probabilities in Interpretations 1 and 2 and the conditional probabilities in Interpretations 3 and 4. The reference point for the unconditional probabilities is the entire group of interest (all hospitalized patients suffering cardiac arrest), whereas the conditional probabilities are interpreted in a more restricted context defined by the "given" event.

Example 6.14 Outcomes of Searches in Traffic Stops

Understand the context ➤ Example 6.11 gave data on the race of drivers in traffic stops in 2019 in Sacramento, California, and whether or not the traffic stop resulted in a search of the driver or vehicle. The Justice Navigator website (referenced in Example 6.11) also gives data on the outcome of searches conducted. The accompanying table gives information on the race of the driver searched and whether or not the search resulted in contraband (such as weapons, drugs, or stolen goods) being found.

Consider the data ➤

		Search Outcome		
		Contraband Found	**Contraband Not Found**	**Total**
Race of Driver	**Black**	663	2159	**2822**
	Latinx	286	850	**1136**
	White	227	573	**800**
	Total	**1176**	**3582**	**4758**

Do the work ➤ Suppose that a traffic stop resulting in a search is randomly selected from these 4758 traffic stops that resulted in searches. Consider the following events:

C = contraband was found
B = driver stopped was Black
L = driver stopped was Latinx
W = driver stopped was White

Copyright 2025 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s). Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

Using the information in the table, you can calculate the following probabilities:

$$P(C) = \frac{1176}{4758} = 0.25$$

$$P(B) = \frac{2822}{4758} = 0.59$$

$$P(L) = \frac{1136}{4758} = 0.24$$

$$P(W) = \frac{800}{4758} = 0.17$$

$$P(C \cap B) = \frac{663}{4758} = 0.14$$

$$P(C \cap L) = \frac{286}{4758} = 0.06$$

$$P(C \cap W) = \frac{227}{4758} = 0.05$$

Now suppose that you know that the driver stopped in the search was Black. How likely is it that the search resulted in contraband being found? To answer this question, calculate the probability $P(C|B)$:

$$P(C \mid B) = \frac{P(C \cap B)}{P(B)} = \frac{0.14}{0.59} = 0.24$$

The probability of contraband being found given that the driver in the search was Latinx is

$$P(C \mid L) = \frac{P(C \cap L)}{P(L)} = \frac{0.06}{0.24} = 0.25$$

The probability of contraband being found given that the driver in the search was White is

$$P(C \mid W) = \frac{P(C \cap W)}{P(W)} = \frac{0.05}{0.17} = 0.29$$

What Do You Think?

How does the conditional probability $P(C \mid B)$ compare to the unconditional probability $P(C)$? What does this difference in these two probabilities reveal?

How does the conditional probability of $P(C \mid L)$ compare to the unconditional probability $P(C)$? What does this difference in these two probabilities reveal?

How does the conditional probability of $P(C \mid W)$ compare to the unconditional probability $P(C)$? What does this difference in these two probabilities reveal?

Interpret the results ➤

Notice that the probability of finding contraband in a search of a Black driver is less than the unconditional probability of finding contraband and the probability of finding contraband in a search of a White driver is greater than the unconditional probability of finding contraband. Even though there were a greater number of searches of Black drivers that found contraband (663) than the number of searches of White drivers that found contraband (227), this was a consequence of the fact that Black drivers were searched in greater numbers than were White drivers. The conditional probabilities reveal that a search of a Black driver was less likely to result in contraband being found compared to the likelihood for White drivers.

Also notice that while the probability of finding contraband in the search of a Latinx driver is equal to the unconditional probability of finding contraband, the probability of that a search will result in contraband being found is still smaller for Latinx drivers that for White drivers. This suggests that the targeting of Black or Latinx drivers for vehicle searches is not justified.

Copyright 2025 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s). Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

As mentioned in the opening paragraphs of this section, one of the most important practical uses of conditional probability is in making diagnoses. A mechanic diagnoses a car by hooking it up to a machine and reading the pressures and speeds of the various components. Doctors observe characteristics of their patients in an attempt to determine whether or not their patients have a certain disease. Many diseases are not actually observable—or at least not easily observable—and often the doctor must make a probabilistic judgment. As illustrated in the next example, conditional probability plays a large role in evaluating diagnostic techniques.

Example 6.15 Diagnosing Tuberculosis

Understand the context ➤

To illustrate the calculations involved in evaluating a diagnostic test, consider the case of tuberculosis (TB), an infectious disease that typically attacks lung tissue. Before 1998, culturing was the standard method for diagnosing TB. This method always resulted in a correct diagnosis, but it took 10 to 15 days to obtain the result. In 1998, investigators evaluated a DNA technique that turned out to be much faster **("LCx: A Diagnostic Alternative for the Early Detection of *Mycobacterium tuberculosis* Complex," *Diagnostic Microbiology and Infectious Diseases* [1998]: 259–264)**.

The DNA technique for detecting tuberculosis was evaluated by comparing results from the test to the existing gold standard (culturing), with the following results for 207 patients exhibiting symptoms:

Consider the data ➤

	Gold Standard Test Shows Has Tuberculosis	Gold Standard Test Shows Does Not Have Tuberculosis
DNA Positive Indication	14	0
DNA Negative Indication	12	181

Converting these data to proportions and adding the column and row totals into the table results in the following information:

Do the work ➤

	Gold Standard Test Shows Has Tuberculosis	Gold Standard Test Shows Does Not Have Tuberculosis	Total
DNA Positive Indication	0.068	0.000	**0.068**
DNA Negative Indication	0.058	0.874	**0.932**
Total	**0.126**	**0.874**	**1.000**

A quick look at the table indicates that the DNA technique seems to be effective as a diagnostic test. Patients who tested positive with the DNA test were also positive with the gold standard test in every case. Patients who tested negative with the DNA test generally tested negative with the gold standard test, but the table also indicates some false-negative results for the DNA test.

Now think about a randomly selected individual who is tested for TB. Consider the following events:

T = event that the individual has tuberculosis
N = event that the DNA test is negative

What is the probability that a person has tuberculosis, given that the DNA test was negative? To answer the question, calculate $P(T|N)$, the probability of the event T given that the event N has occurred. You can calculate this probability as follows:

$$P(T|N) = P(\text{tuberculosis}|\text{negative DNA test})$$

$$= \frac{P(\text{tuberculosis} \cap \text{negative DNA test})}{P(\text{negative DNA test})}$$

$$= \frac{0.058}{0.932}$$
$$= 0.062$$

Interpret the results ➤ Notice that 0.126 is the proportion of those tested who had tuberculosis. The added information provided by the diagnostic test has altered probability—and provides some measure of relief for patients who test negative. Once it is known that the test result is negative, the estimated probability of having tuberculosis is only about half as large.

Exercises 6.38–6.52

6.38 The article **"Can Halloween Costumes Influence a Trick-or-Treater's Honesty" (*The Los Angeles Times*, October 31, 2019)** describes an experiment to test whether type of costume is related to trick-or-treaters' tendency to cheat. The researchers conducting the study suspected that kids wearing "good guy" costumes would be more honest than kids wearing "bad guy" costumes. They asked 544 trick-or-treaters whether they were dressed as a good guy or a bad guy and then had them play a game that offered an easy opportunity to cheat to determine whether they would get one or two pieces of candy. About 74% of the trick-or-treaters said they were good guys and about 24% said that they were bad guys (the remaining 2% said they were both or neither). About 28% of kids who said they were wearing a good guy costumes cheated to get an extra candy and about 51% of the kids who said they were wearing a bad guy costume cheated. Five percentages are given. Each one represents a probability associated with the chance experiment of randomly selecting one of these trick-or-treaters. Which of these percentages represents conditional probabilities?

6.39 The authors of the paper **"Racial-Ethnic Disparities in Opioid Prescriptions at Emergency Department Visits for Conditions Commonly Associated with Prescription Drug Abuse" (*PLOS ONE*, August 8, 2016)** investigated racial differences in the medications offered for similar medical conditions. They used data from a national sample of visits to emergency rooms over a 5-year period. For patients visiting an emergency department for severe abdominal pain, data given in the paper were used to estimate the following probabilities:

P(patient is given a prescription for an opioid pain medication) $= 0.52$
P(patient is non-Hispanic Black) $= 0.23$
P(patient is non-Hispanic Black and is given a prescription for an opioid pain medication) $= 0.08$

a. Calculate the probability that a patient is given a prescription for an opioid pain medication given that the patient is non-Hispanic Black.

b. How does the conditional probability calculated in Part (a) compare to the unconditional probability that a patient visiting the emergency department for severe abdominal pain is given a prescription for an opioid pain medication?

c. The same type of difference you found in Part (b) for non-Hispanic Black patients visiting the emergency department for severe abdominal pain was also noted for Hispanic patients visiting the emergency department for severe abdominal pain and also for non-Hispanic Black and Hispanic patients visiting the emergency department for back pain. How does this support the authors conclusion that there are racial disparities in the treatment of abdominal pain and back pain?

6.40 The 2020 Consumer Financial Literacy Survey summarized data from a representative sample of adult Americans. Survey participants were asked if they felt confident in their most recent big financial decision and were also classified into groups by age. Data consistent with summary percentages given in the report are summarized in the accompanying table.

	Confident in Decision	**Not Confident in Decision**
Age 18to 34	260	50
Age 35 to 44	141	29
Age 45 to 54	174	26
Age 55 to 64	160	20
Age 65 and Older	133	7

Suppose that one of these survey participants is selected at random and consider the following events:

C = event that the selected person is confident in their most recent financial decision
Y = event that the selected person is 44 years old or younger
O = event that the selected person is 45 years old or older

a. Use the data in the given table to estimate the following probabilities.
   i. $P(C)$
   ii. $P(C|Y)$
   iii. $P(C|O)$

b. Interpret each of the probabilities calculated in Part (a).

6.41 Ipsos, a market research company, surveyed a representative sample of 1005 adult Americans. Each person in the sample was asked about the region of the country where they lived and whether they had a tattoo or multiple tattoos on their body **("More Americans Have Tattoos Today Than Seven Years Ago," Ipsos.com, August 29, 2019)**. The responses are summarized in the accompanying table.

	Region			
	Northeast	**Midwest**	**South**	**West**
One Tattoo	27	33	42	30
Multiple Tattoos	27	36	66	41
No Tattoos	125	143	269	166

Suppose that one of these survey participants is selected at random and consider the following events:

W = event that the selected person lived in the West

N = event that the selected person lived in the Northeast

T = event that the selected person has at least one tattoo

a. Use the data in the given table to estimate the following probabilities:
   i. $P(T)$
   ii. $P(T|W)$
   iii. $P(T|N)$

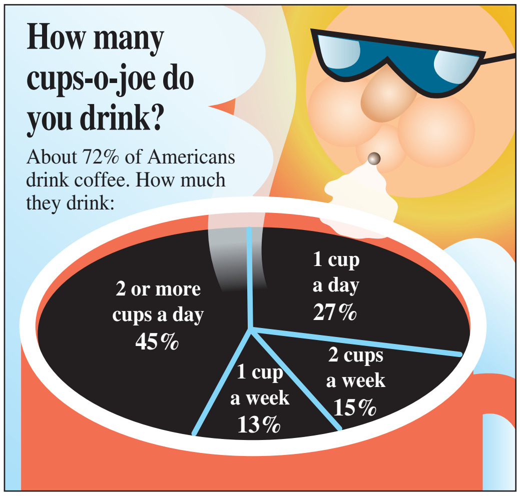
b. Interpret each of the probabilities calculated in Part (a).

c. How do the probabilities calculated in Part (a) compare and what does this tell you about possible regional differences?

6.42 The following graphical display is similar to one that appeared in ***USA TODAY* (January 8, 2010)**. Use the information in this display to answer the following questions. Assume that the percentages in the graph are representative of adult Americans.

a. What is the probability that a randomly selected adult American drinks coffee?

b. The display associates 45% with the category "two or more cups a day." For the chance experiment that consists of selecting an adult American at random, is 0.45 the probability that the selected individual drinks two or more cups of coffee a day, or is it the conditional probability that the selected individual drinks two or more cups a day given that they drink coffee? Explain.



Source: Braun Research for International Delight Coffee House Inspirations; USA TODAY, January 8, 2010.

6.43 The article **"Americans Growing More Concerned About Head Injuries in Football" (theharrispoll.com, December 21, 2015)** describes a survey of 2096 adult Americans. Survey participants were asked if they were football fans and also if they agreed or disagreed that the rules that the National Football League adopted in 2010 designed to limit head injuries have been effective. Data from the survey are summarized in the table below.

	Agree	**Disagree**	**Total**
Football Fan	693	523	1,216
Not a Football Fan	229	651	880
Total	922	1,174	2,096

Suppose that a survey participant is to be selected at random. Consider the following events:

A = event selected participant agrees that the rules have been effective

D = event selected participant disagrees that the rules have been effective

F = event selected participant is a football fan

Calculate the following probabilities

a. $P(A)$

b. $P(D)$

c. $P(A|F)$

d. $P(A|F^C)$

6.44 For each of the probabilities calculated in the previous exercise, write a sentence that provides an interpretation of that probability.

6.45 The accompanying data are from the ***Chronicle for Higher Education Almanac* (2021)**. Each person in a representative sample of 3337 students at two-year colleges was asked how they would describe their household's financial situation compared to before the COVID-19 pandemic. Each person in the sample was also asked if they have children who depend on them for care living with them. The responses are summarized in the given table.

Copyright 2025 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s). Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

	Better than Before	**Same as Before**	**Worse than Before**
Children	40	299	310
No Children	210	1512	966

Suppose that one of these 3337 two-year college students is to be selected at random.

a. What is the probability that the selected student has children?

b. What is the probability that the selected student responded that they are better off than before the pandemic?

c. What is the probability that the selected student responded that they are better off than before the pandemic, given that they have children?

d. What is the probability that the selected student responded that they are worse off than before the pandemic, given that they have children?

6.46 Suppose that an individual is randomly selected from the population of all adult males living in the United States. Let A be the event that the selected individual is over 6 feet in height, and let B be the event that the selected individual is a professional basketball player. Which do you think is larger, $P(A|B)$ or $P(B|A)$? Why?

6.47 Is ultrasound a reliable method for determining the biological sex of an unborn baby? The accompanying data on 1000 births are consistent with summary values that appeared in the ***Journal of Statistics Education* ("New Approaches to Learning Probability in the First Statistics Course," 2001)**.

	Ultrasound Predicted Sex: Female	**Ultrasound Predicted Sex: Male**
Sex assigned at birth: Female	432	48
Sex assigned at birth: Male	130	390

a. Use the given information to estimate the probability that a newborn baby is assigned female at birth, given that the ultrasound predicted the baby would be female.

b. Use the given information to estimate the probability that a newborn baby is assigned male at birth, given that the ultrasound predicted the baby would be male.

c. Based on your answers to Parts (a) and (b), do you think that an ultrasound prediction of male and an ultrasound prediction of female are equally reliable? Explain.

6.48 As of February 2023, 10 U.S. states have introduced legislation banning the sale of flavored tobacco products, and the Food and Drug Administration has proposed a ban on the sale of all flavored cigars. The authors of the paper **"Association of Flavor Perception with Blue vs Purple Cigar Packaging Among U.S. Adults" (*JAMA Network Open* [2023])** investigated whether the color of the packaging conveyed a message about whether a cigar was flavored. Adults were shown a cigar in either blue or purple packaging and then were asked if they thought that the cigar was flavored. Data from this experiment is summarized in the accompanying table.

	Response		
	Flavored	**Not Flavored**	**Total**
Blue Packaging	670	649	1319
Purple Packaging	846	376	1222
Total	1516	1025	2541

a. How many adults participated in the study? How many of the participants viewed the blue packaging?

b. Suppose that one of these study participants is selected at random. Calculate the following probabilities.

  i. The probability that the selected participant viewed the blue packaging.

  ii. The probability that the selected participant responded that they thought the cigar was flavored.

  iii. The probability that the selected participant thought that the cigar was flavored given that they viewed the blue packaging.

  iv. The probability that the selected participant thought that the cigar was flavored given that they viewed the purple packaging.

c. The paper states that the study results "suggested that purple communicated the product was flavored more so than blue." Provide a justification for this conclusion based on the probabilities that you calculated in Part (b).

6.49 The report **"2015 Utah Seat Belt Use Survey" (Utah Department of Public Safety—Highway Safety Office, September 14, 2015)** stated that based on observing a large number of vehicle occupants, the estimated percentage of Utah drivers and passengers who wear seatbelts is 87.2%. The report also gave information on seat belt use and whether the vehicle is traveling in an urban or rural area. The information in the following table is consistent with summary values given in the report.

	Urban Areas	**Rural Areas**
Seat Belt	1799	1607
No Seat Belt	201	393

Copyright 2025 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s). Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

Assume that these data are representative of drivers and passengers in Utah. Consider the following conclusion:

> Drivers in urban areas are more likely to wear seat belts than drivers in rural areas.

Provide a justification for this conclusion. Use the information in the table to calculate relevant probabilities.

6.50 Contract cheating is a form of plagiarism in which students pay someone else to complete an assignment for them. The authors of the paper **"How Prevalent is Contract Cheating and to What Extent are Students Repeat Offenders" (*Journal of Academic Ethics* [2017]: 115–124)** collected data from college students by presenting them with the following scenario: "A student has to write an essay for their economics class. They find a website on the Internet where they can pay someone to write it for them. They pay the money and are emailed the essay. They hand the essay in as their own." Students are then asked if they had done a similar thing. Students were also classified according to whether they were a first-year college student. The following statements are based on the study results:

1. 46.5% of the study participants were first-year students.
2. 2.1% of the study participants reported contract cheating.
3. 36.4% of the students who reported they had engaged in contract cheating were first-year students.
4. 46.7% of the students who reported they had not engaged in contract cheating were first-year students.

Each of the four given percentages can be interpreted as a probability. Which of these probabilities are conditional probabilities and which are not conditional probabilities?

6.51 The article **"U.S. Investors Split Between Digital and Traditional Banking" (www.gallup.com, August 5, 2016)** summarized data from a Gallup survey of a random sample of 1019 American adults with investments of $10,000 or more. Based on the survey data, it was estimated that 31% of investors manage their investments by doing everything they possibly can online. But the authors of the article also noted that there was a quite a difference between younger investors (ages 18–49) and older investors (ages 50 and older). For younger investors, 43% said they do everything they possibly can online, while the percentage for older investors was 23%.

Consider the following events:

O = event that a randomly selected investor does everything possible online

A = event that a randomly selected investor is age 18 to 49

B = event that a randomly selected investor is age 50 or older

Use the given information to estimate the following probabilities:

a. $P(O)$
b. $P(O|A)$
c. $P(O|B)$

6.52 The Bureau of Transportation website provides data on the on-time performance of airlines. The tables below summarize data from this website for flights that occurred between July 2020 and July 2021.

a. The following table summarizes data for flights leaving the Phoenix and Seattle airports for two airlines, American and Alaska Airlines.

	On-Time Departure	Late Departure	Total
American Airlines	32,996	4,545	**37,541**
Alaska Airlines	46,777	6,526	**53,303**
Total	**79,773**	**11,071**	**90,844**

Suppose that one of these flights is selected at random.

i. What is the probability that the flight had an on-time departure?
ii. What is the probability that the flight had an on-time departure given that it was an American Airlines flight?
iii. What is the probability that the flight had an on-time departure given that it was an Alaska Airlines flight?
iv. Which airline had the better on-time performance?

b. Now consider the flight data for just the Phoenix airport.

Phoenix Airport	On-Time Departure	Late Departure	Total
American Airlines	29,397	3,782	**33,179**
Alaska Airlines	2,784	299	**3,083**
Total	**32,181**	**4,081**	**36,262**

Suppose that one of these flights from the Phoenix airport is selected at random.

i. What is the probability that the flight had an on-time departure?
ii. What is the probability that the flight had an on-time departure given that it was an American Airlines flight?
iii. What is the probability that the flight had an on-time departure given that it was an Alaska Airlines flight?
iv. Which airline had the better on-time performance?

c. Now consider the flight data for just the Seattle airport.

Copyright 2025 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s). Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

Seattle Airport	**On-Time Departure**	**Late Departure**	**Total**
American Airlines	3,599	763	**4,362**
Alaska Airlines	43,993	6,227	**50,220**
Total	**47,592**	**6,990**	**54,582**

Suppose that one of these flights from the Seattle airport is selected at random.

i. What is the probability that the flight had an on-time departure?

ii. What is the probability that the flight had an on-time departure given that it was an American Airlines flight?

iii. What is the probability that the flight had an on-time departure given that it was an Alaska Airlines flight?

iv. Which airline had the better on-time performance?

d. You should have noticed from Parts (b) and (c) that for both the Phoenix airport and the Seattle airport, Alaska Airlines had the better on-time performance. But in Part (a) when the data for both airports is combined, it looks like American has slightly better on-time performance. This is an example of what is known as Simpson's Paradox. Write a brief explanation of why the apparent inconsistency occurs for this data set. (Hint: Does it look like the airports differ with respect to flight delays?)

Section 6.5 Independence

In Section 6.4, you saw that knowing one event has occurred can change the assessment of the probability that some other event has occurred. However, it is also possible that knowing that one event has occurred will not change the assessment of the probability of occurrence of a second event.

Example 6.16 Mortgage Choices

Understand the context ➤

A bank offers both adjustable-rate and fixed-rate mortgage loans on residential property, which it classifies into three categories: single-family houses, condominiums, and multifamily dwellings. The following table, called a *joint probability table*, displays probabilities based on the bank's long-run lending behavior:

Consider the data ➤

	Single-Family	**Condo**	**Multifamily**	**Total**
Adjustable	0.40	0.21	0.09	**0.70**
Fixed	0.10	0.09	0.11	**0.30**
Total	**0.50**	**0.30**	**0.20**	**1.00**

From the table you can see that 70% of all mortgages are adjustable rate, 50% of all mortgages are for single-family houses, 40% of all mortgages are adjustable rate for single-family houses (adjustable-rate *and* single-family), and so on.

Define the events E and F as

E = event that a mortgage is adjustable rate
F = event that a mortgage is for a single-family house

Do the work ➤

Then $P(E) = 0.70$ and

$$P(E|F) = \frac{P(E \text{ and } F)}{P(F)} = \frac{0.40}{0.50} = 0.80$$

Interpret the results ➤

This means that 80% of loans made for single-family houses are adjustable-rate loans. Notice that $P(E|F)$ is larger than the original (unconditional) probability $P(E) = 0.70$. Also,

$$P(F|E) = \frac{P(E \text{ and } F)}{P(E)} = \frac{0.40}{0.70} = 0.571$$

which is larger than the unconditional probability $P(F) = 0.5$. Knowing that E (adjustable rate) has occurred has changed the assessment of how likely it is that F (single-family house) has also occurred.

Now consider another event C defined as

C = event that a mortgage is for a condominium

Copyright 2025 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s). Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

then

$$P(E|C) = \frac{P(E \text{ and } C)}{P(C)} = \frac{0.21}{0.30} = 0.70$$

Notice that $P(E|C) = P(E)$. In this case, knowing that a mortgage is for a condominium doesn't change the assessment of the probability that the mortgage has an adjustable interest rate.

When two events E and F are such that $P(E|F) = P(E)$, the probability that event E has occurred is the same after you learn that F has occurred as it was before you knew that F had occurred. If this is the case, you would say that E and F are independent of one another.

Definitions

Independent events: Two events E and F are **independent** if

$$P(E|F) = P(E)$$

Dependent events: If two events E and F are not independent, they are **dependent** events.

If $P(E|F) = P(E)$, it is also true that $P(F|E) = P(F)$, and vice versa.

Independence of events E and F also implies the following additional three relationships:

$$P(\text{not } E|F) = P(\text{not } E)$$
$$P(E|\text{not } F) = P(E)$$
$$P(\text{not } E|\text{not } F) = P(\text{not } E)$$

This means that if E and F are independent, nothing you learn about F will change the probability of E or of *not* E.

Recall that the formula for conditional probability is

$$P(E|F) = \frac{P(E \cap F)}{P(F)}$$

which can be rearranged to give

$$P(E \cap F) = P(E|F)P(F)$$

When E and F are independent, $P(E|F) = P(E)$, so it follows that if E and F are independent,

$$P(E \cap F) = P(E|F)P(F) = P(E)P(F)$$

This result is called the multiplication rule for two independent events.

Multiplication Rule for Two Independent Events

If the events E and F are independent,

$$P(E \cap F) = P(E)P(F)$$

Example 6.17 Hitchhiker's Thumb

Understand the context ➤

In humans, there is a gene that controls a characteristic known as hitchhiker's thumb. Hitchhiker's thumb is the ability to bend the last joint of the thumb back at an angle of 60° or more. Whether a child has hitchhiker's thumb is determined by two random events:

Copyright 2025 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s). Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

which of two alleles is contributed by the father and which of two alleles is contributed by the mother. You can think of these alleles as a parental vote of yes or no on the hitchhiker's thumb gene. If the votes by the two parents disagree, the dominant allele wins and the child does not have hitchhiker's thumb. These two random events, the results of cell division in two different biological parents, are independent of each other.

Suppose that there is a 0.10 probability that a parent contributes a positive hitchhiker's thumb allele. Because the events are independent, the probability that both parents contribute a positive hitchhiker's thumb allele, H+, to the offspring is

Do the work ➤

$$\begin{aligned} &P(\textit{mother contributes } H+ \cap \textit{father contributes } H+) \\ &\quad = P(\textit{mother contributes } H+)P(\textit{father contributes } H+) \\ &\quad = (0.10)(0.10) \\ &\quad = 0.01 \end{aligned}$$

Interpret the results ➤

This means that only about 1% of all children in the population would have hitchhiker's thumb.

Example 6.18 Curious Guppies

Now consider another example, this time from the field of animal behavior, to illustrate how an investigator could judge whether two events are independent. In a number of fish species, including guppies, a phenomenon known as predator inspection has been reported. It is thought that predator inspection allows a guppy to assess the risk posed by a potential predator. In a typical inspection a guppy moves toward a predator, presumably to acquire information and then (hopefully) depart to inspect again another day.

Understand the context ➤

Investigators have observed that guppies sometimes approach and inspect a predator in pairs. Suppose that it is not known whether these predator inspections are independent or whether the guppies are operating as a team. Let p denote the probability that an individual guppy will inspect a predator. Suppose E_1 is the event that guppy 1 will approach and inspect a predator and E_2 is the event that guppy 2 will approach and inspect a predator. Then the probability of guppies 1 and 2 both approaching the predator at the same time by chance *if* they are acting *independently* is

$$P(E_1 \cap E_2) = P(E_1)P(E_2) = p \cdot p = p^2$$

Interpret the results ➤

Based on this analysis, if the inspections are in fact independent, you would expect the proportion of times that two guppies happen to simultaneously inspect a predator to be equal to the square of the proportion of times that a single fish does so. For example, if the probability a single guppy inspects a predator is 0.3, you would expect the proportion of the time that two guppies would inspect a predator simultaneously to be $(0.3)^2 = 0.09$. Based on observations of the inspection behavior of a large number of guppies, scientists found that inspections by two guppies occurred much more often than 9% of the time. As a result, they concluded that the inspection behavior of guppies does not appear to be independent.

The concept of independence extends to more than two events. Consider three events, E_1, E_2, and E_3. Then independence means not only that

$$\begin{aligned} P(E_1|E_2) &= P(E_1) \\ P(E_3|E_2) &= P(E_3) \end{aligned}$$

and so on but also that

$$\begin{aligned} P(E_1|E_2 \textit{ and } E_3) &= P(E_1) \\ P(E_1 \textit{ and } E_3|E_2) &= P(E_1 \textit{ and } E_3) \end{aligned}$$

and so on. There is also a multiplication rule for more than two independent events.

Multiplication Rule for *k* Independent Events

Events $E_1, E_2, \ldots, E_k$ are **independent** if knowledge that any of the events have occurred does not change the probabilities that any particular one or more of the other events has occurred.

Independence implies that

$$P(E_1 \cap E_2 \cap \cdots \cap E_k) = P(E_1)P(E_2) \cdots P(E_k)$$

This means that when events are independent, the probability that all occur together is the product of the individual probabilities. This relationship also holds if one or more of the events is replaced by its complement.

Example 6.19 takes a rather simplified view of a car to illustrate the use of the multiplication rule.

Example 6.19 Car Configurations

Understand the context ➤

A new car has four critical components that all need to function properly in order for the car to work: the engine, the fuel system, the ignition system, and the cooling system.

Suppose that for a newly assembled car

E_1 = event that the engine functions properly
E_2 = event that the fuel system functions properly
E_3 = event that the ignition system functions properly
E_4 = event that the cooling system functions properly

Suppose these four events are independent, with

$$P(E_1) = P(E_2) = 0.98 \qquad P(E_3) = 0.95 \qquad P(E_4) = 0.99$$

Do the work ➤

The probability that all these systems function properly is then

$$\begin{aligned} P(E_1 \cap E_2 \cap E_3 \cap E_4) &= P(E_1)P(E_2)P(E_3)P(E_4) \\ &= (0.98)(0.98)(0.94)(0.99) \\ &= 0.89 \end{aligned}$$

Interpret the result ➤

This probability can be interpreted as follows: In the long run, 89% of newly assembled cars would have all systems functioning properly. (In reality, the reliability of these systems is much higher than the numbers used in this example!) The probability that all systems except the engine will function properly is

$$\begin{aligned} P(E_1^C \cap E_2 \cap E_3 \cap E_4) &= P(E_1^C)P(E_2)P(E_3)P(E_4) \\ &= (1 - P(E_1))P(E_2)P(E_3)P(E_4) \\ &= (0.02)(0.98)(0.94)(0.99) \\ &= 0.018 \end{aligned}$$

Sampling With and Without Replacement

One area of statistics where the rules of probability are important is sampling. As you saw in Chapter 2, a well-designed sampling plan allows investigators to make inferences about a population based on information from a sample. Sampling methods can be classified into two categories: **sampling with replacement** and **sampling without replacement**.

Most inferential methods presented in an introductory statistics course are based on the assumption of sampling *with* replacement, but when sampling from real populations, you almost always sample *without* replacement. This seemingly contradictory practice can be a source of confusion. Fortunately, under certain conditions, the distinction between sampling with and without replacement is not important.

Copyright 2025 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s). Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

Definitions

Sampling with replacement: Once selected, an individual or object is put back into the population before the next selection.

Sampling without replacement: Once selected, an individual or object is not returned to the population prior to subsequent selections.

Example 6.20 Sampling With and Without Replacement

Consider the process of selecting three cards from a standard deck of playing cards. A standard deck of cards contains cards of four types (clubs, diamonds, hearts, and spades), with 13 cards of each type. This selection can be made in two ways. One method is to shuffle the cards and then deal three cards off the top of the deck. This would be sampling without replacement. A second method, rarely seen in real games, is to select a card at random, note which card is observed, replace it in the deck, and shuffle before selecting the next card. This method is sampling with replacement.

From the standpoint of probability, sampling with and without replacement are analyzed differently. To see this, consider these events:

H_1 = event that the first card is a heart
H_2 = event that the second card is a heart
H_3 = event that the third card is a heart

Because 13 of the 52 cards are hearts, if sampling is done with replacement, the probability of H_3 is 0.25, regardless of whether either H_1 or H_2 occurs, because replacing selected cards gives the same deck for the third selection as for the first two selections. Whether either of the first two cards was a heart has no bearing on the third card selected, so the three events, H_1, H_2, and H_3, are independent.

When sampling is without replacement, the chance of getting a heart on the third card selected does depend on the results of the first two selections. If both H_1 and H_2 occur, only 11 of the 50 remaining cards are hearts. Because any one of these 50 has the same chance of being selected, the probability of H_3 in this case is

$$P(H_3|H_1 \text{ and } H_2) = \frac{11}{50} = 0.22$$

Alternatively, if neither of the first 2 cards is a heart, then all 13 hearts remain in the deck for the third draw, so

$$P(H_3| \text{ not } H_1 \text{ and not } H_2) = \frac{13}{50} = 0.26$$

Information about the occurrence of H_1 and H_2 affects the chance that H_3 has occurred. For sampling without replacement, the three events described here are not independent.

In opinion polls and other types of surveys, sampling is almost always done without replacement. For this method of sampling, the results of successive selections are not independent of one another. However, Example 6.21 suggests that, *under certain conditions*, the fact that selections in sampling without replacement are not independent is not a cause for concern.

Example 6.21 Independence and Sampling Without Replacement

Suppose that a shipment of 10,000 circuit boards used in graphing calculators consists of 2500 manufactured by one firm and 7500 manufactured by a second firm, all mixed together. Three of the chips will be selected at random *without* replacement. Let

E_1 = event that first circuit board selected was manufactured by Firm 1
E_2 = event that second circuit board selected was manufactured by Firm 1
E_3 = event that third circuit board selected was manufactured by Firm 1

Following the reasoning used in Example 6.20, you can calculate

$$P(E_3|E_1 \text{ and } E_2) = \frac{2498}{9998} = 0.24985$$

$$P(E_3|\text{not } E_1 \text{ and not } E_2) = \frac{2500}{9998} = 0.25005$$

Although these two probabilities differ slightly, when rounded to three decimal places they are both 0.250. The occurrence or nonoccurrence of E_1 or E_2 has virtually no effect on the chance that E_3 will occur. *For practical purposes*, the three events can be considered independent.

The essential difference between the situations described in Example 6.20 and Example 6.21 is the size of the sample relative to the size of the population. In Example 6.20 a relatively large proportion of the population was sampled (3 out of 52), whereas in Example 6.21 the proportion of the population sampled was quite small (only 3 out of 10,000).

Usually, when a sample is selected, the sample size is small compared to the size of the population. The condition of sampling with replacement required for many inferential methods can coexist with the practice of sampling without replacement because of the following principle:

> If a random sample of size n is selected from a population of size N, the probabilities of successive selections calculated on the basis of sampling with replacement and on the basis of sampling without replacement are nearly equal when n is small compared to N.
>
> In practice, independence can be assumed for the purpose of calculating probabilities as long as n is not larger than 5% of N.

This principle justifies the assumption of independence in many statistical problems. The phrase *assumption of independence* does not signify that the investigators are in some sense fooling themselves. They are recognizing that, for practical purposes, the results will not differ from the "right" answers.

In some sampling situations, the sample size might be more than 5% of the population. For example, a newspaper editor at a small school might easily sample more than 5% of the students at the school to assess student opinion on a particular issue. In that case, the editors would be wise to consult a statistician before proceeding. Does this mean that an investigator should never sample more than 5% of a population? Certainly not! It is almost always the case that a larger sample results in better inferences about the population. The only disadvantage of sampling more than 5% of the population (other than an increase in the time and resources required) is that the analysis of the resulting data is slightly more complicated.

Exercises 6.53–6.68

6.53 Consider a chance experiment that consists of selecting a student at random from a high school with 3000 students.

a. In the context of this chance experiment, give an example of two events that would be mutually exclusive.

b. In the context of this chance experiment, give an example of two events that would not be mutually exclusive.

6.54 In 2019, IPSOS surveyed people in a representative sample of American adults to determine how common tattoos are **("More Americans Have Tattoos Today Than Seven Years Ago," August 29, 2019)**. They found that 24% of college graduates and 33% of those without a college degree had at least one tattoo. Suppose that this information is representative of American adults. If an American adult is selected at random, are the events

Copyright 2025 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s). Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

selected adult has a college degree and *selected adult has at least one tattoo* independent or dependent? Explain.

6.55 Many fire stations handle emergency calls for medical assistance as well as calls requesting firefighting equipment. A particular station says that the probability that an incoming call is for medical assistance is 0.85. This can be expressed as $P(\textit{call is for medical assistance}) = 0.85$.

a. Give a relative frequency interpretation of the given probability.
b. What is the probability that a call is not for medical assistance?

6.56 Refer to the information given in the previous exercise.

a. Assuming that successive calls are independent of one another, calculate the probability that two successive calls will both be for medical assistance. (Hint: See Example 6.18.)
b. Still assuming independence, calculate the probability that for two successive calls, the first is for medical assistance and the second is not for medical assistance.
c. Still assuming independence, calculate the probability that exactly one of the next two calls will be for medical assistance. (Hint: There are two different possibilities. The one call for medical assistance might be the first call, or it might be the second call.)
d. Do you think that it is reasonable to assume that the requests made in successive calls are independent? Explain.

6.57 The report **"Great Jobs, Great Lives. The Relationship Between Student Debt, Experiences and Perceptions of College Worth" (Gallup-Purdue Index 2015 Report)** gave information on the percentage of recent college graduates (those graduating between 2006 and 2015, inclusive) who strongly agree with the statement "My college education was worth the cost." Suppose that a college graduate will be selected at random, and consider the following events:

A = event that the selected graduate strongly agrees that education was worth the cost
N = event that the selected graduate finished college with no student debt
H = event that the selected graduate finished college with high student debt (over \$50,000)

The following probability estimates were given in the report:

$$P(A) = 0.38 \qquad P(A|N) = 0.49 \qquad P(A|H) = 0.18$$

a. Interpret the value of $P(A|N)$.
b. Interpret the value of $P(A|H)$.
c. Are the events A and H independent? Justify your answer.

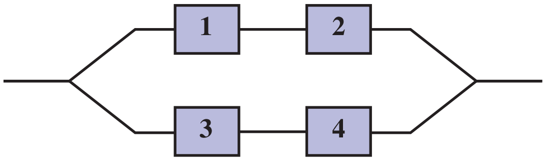
6.58 In a small city, approximately 15% of those eligible are called for jury duty in any one calendar year. People are selected for jury duty at random from those eligible, and the same individual cannot be called more than once in the same year.

a. What is the probability that a particular eligible person in this city is selected in both of the next 2 years?
b. What is the probability that a particular eligible person in this city is selected in all three of the next 3 years?

6.59 Jordan is a bit forgetful, and if they don't make a "to do" list, the probability that Jordan forgets something they are supposed to do is 0.1. Tomorrow Jordan intends to run three errands, and fails to write them on their list.

a. What is the probability that Jordan forgets all three errands? What assumptions did you make to calculate this probability?
b. What is the probability that Jordan remembers at least one of the three errands?
c. What is the probability that Jordan remembers the first errand but not the second or third?

6.60 Consider a system consisting of four components, as pictured in the following diagram:



Components 1 and 2 form a series subsystem, as do Components 3 and 4. The two subsystems are connected in parallel. Suppose that

$P(1 \text{ works}) = 0.9$
$P(2 \text{ works}) = 0.9$
$P(3 \text{ works}) = 0.9$
$P(4 \text{ works}) = 0.9$

and that the four components work independently of one another.

a. The 1–2 subsystem works only if both components work. What is the probability of this happening?
b. What is the probability that the 1–2 subsystem doesn't work? that the 3–4 subsystem doesn't work?
c. The system won't work if the 1–2 subsystem doesn't work and if the 3–4 subsystem also doesn't work. What is the probability that the system won't work? that it will work?

Copyright 2025 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s). Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

6.61 Consider the system described in the previous exercise.

- **a.** How would the probability of the system working change if a 5–6 subsystem were added in parallel with the other two subsystems?
- **b.** How would the probability that the system works change if there were three components in series in each of the two subsystems?

6.62 In a January 2016 Harris Poll, each of 2252 American adults was asked the following question: If you had to choose, which ONE of the following sports would you say is your favorite?" **("Pro Football is Still America's Favorite Sport," theharrispoll.com).** Of the survey participants, 33% chose pro football as their favorite sport. The report also included the following statement "Adults with household incomes of \$75,000 to < \$100,000 (48%) are especially likely to name pro football as their favorite sport, while love of this particular game is especially low among those in \$100,000+ households (21%)."

Suppose that the percentages from this poll are representative of American adults in general. Consider the following events:

F = event that a randomly selected American adult names pro football as their favorite sport
L = event that a randomly selected American has a household income of \$75,000 to < \$100,000
H = event that a randomly selected American has a household income of \$100,000+

- **a.** Use the given information to estimate the following probabilities:
  - **i.** $P(F)$
  - **ii.** $P(F|L)$
  - **iii.** $P(F|H)$
- **b.** Are the events F and L mutually exclusive? Justify your answer.
- **c.** Are the events H and L mutually exclusive? Justify your answer.
- **d.** Are the events F and H independent? Justify your answer.

6.63 Consider the following events:

T = event that a randomly selected adult trusts credit card companies to safeguard their personal data
M = event that a randomly selected adult is between the ages of 19 and 36
O = event that a randomly selected adult is 37 or older

Based on a June 9, 2016, Gallup survey **("Data Security: Not a Big Concern for Millennials," gallup.com)**, the following probability estimates are reasonable:

$$P(T|M) = 0.27 \qquad P(T|O) = 0.22$$

Explain why $P(T)$ is not just the average of the two given probabilities.

6.64 The following case study was reported in the article **"Parking Tickets and Missing Women,"** which appeared in an early edition of the book ***Statistics: A Guide to the Unknown***. In a Swedish trial on a charge of overtime parking, a police officer testified that he had noted the position of the two air valves on the tires of a parked car: To the closest hour, one was at the one o'clock position and the other was at the six o'clock position. After the allowable time for parking in that zone had passed, the policeman returned, noted that the valves were in the same position, and ticketed the car. The owner of the car claimed that they had left the parking place in time and had returned later. The valves just happened by chance to be in the same positions.

An "expert" witness computed the probability of this occurring as $(1/12)(1/12) = 1/144$.

- **a.** What reasoning did the expert use to arrive at the probability of 1/144?
- **b.** Can you spot the error in the reasoning that leads to the stated probability of 1/144?
- **c.** What effect does this error have on the probability of occurrence?
- **d.** Do you think that 1/144 is larger or smaller than the correct probability of occurrence?

6.65 Three friends (A, B, and C) will participate in a round-robin tournament in which each one plays both of the others. Suppose that

$P(\text{A beats B}) = 0.7$
$P(\text{A beats C}) = 0.8$
$P(\text{B beats C}) = 0.6$

and that the outcomes of the three matches are independent of one another.

- **a.** What is the probability that A wins both matches and that B beats C?
- **b.** What is the probability that A wins both matches?
- **c.** What is the probability that A loses both matches?
- **d.** What is the probability that each person wins one match? (Hint: There are two different ways for this to happen.)

6.66 False positive results are not uncommon with mammograms, a test used to screen for breast cancer. For a patient who has a positive mammogram, the probability that they actually have breast cancer is less than 0.05 if they are under 40 years old, and ranges from 0.050 to 0.109 if they are over 40 years old

("Breast Cancer Screenings: Does the Evidence Support the Recommendations?," *Significance* [August 2016]: 24–37). If a patient with a positive mammogram is selected at random, are the two events

B = selected patient has breast cancer

and

A = selected patient is over 40 years old

independent events? Justify your answer using the given information.

6.67 The **National Public Radio** show ***Car Talk*** used to have a feature called "The Puzzler." Listeners were asked to send in answers to some puzzling questions—usually about cars but sometimes about probability (which, of course, must account for the incredible popularity of the program!).

Suppose that for a car question, 800 answers were submitted, of which 50 are correct.

a. Suppose that the hosts randomly select two answers from those submitted *with replacement*. Calculate the probability that both selected answers are correct. (For purposes of this problem, keep at least five digits to the right of the decimal.)

b. Suppose now that the hosts select the answers at random but *without replacement*. Use conditional probability to evaluate the probability that both answers selected are correct. How does this probability compare to the one computed in Part (a)?

6.68 Refer to the previous exercise. Suppose now that for a probability question, 100 answers are submitted, of which 50 are correct. Calculate the probabilities in Parts (a) and (b) of the previous exercise for the probability question.

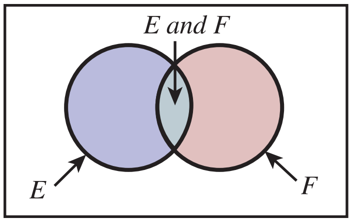
Section 6.6 Some General Probability Rules

In previous sections, you saw how the probability of $P(E \cup F)$ could be easily calculated when E and F are mutually exclusive and how $P(E \cap F)$ could be calculated when E and F are independent. In this section, you will see more general rules: an addition rule that can be used even when events are not mutually exclusive and a multiplication rule that can be used even when events are not independent.

General Addition Rule

Calculating $P(E \cup F)$ when the two events are not mutually exclusive is a bit more complicated than in the case of mutually exclusive events. Consider Figure 6.9, in which E and F overlap. The area of the colored region $(E \cup F)$ is not the sum of the area of E and the area of F, because when the two individual areas are added, the area of the intersection $(E \cap F)$ is counted twice. Similarly, $P(E) + P(F)$ includes $P(E \cap F)$ twice, so this intersection probability must be subtracted from the sum to obtain $P(E \cup F)$. This reasoning leads to the general addition rule.

Figure 6.9
The colored region represents $P(E \cup F)$, and $P(E \cup F) \neq P(E) + P(F)$.



General Addition Rule for Two Events

For any two events E and F,

$$P(E \cup F) = P(E) + P(F) - P(E \cap F)$$

Copyright 2025 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s). Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

When E and F are mutually exclusive, the general addition rule simplifies to the previous rule for mutually exclusive events. This is because when E and F are mutually exclusive, $E \cap F$ contains no outcomes and $P(E \cap F) = 0$. The general addition rule can be used to determine any one of the four probabilities $P(E)$, $P(F)$, $P(E \cap F)$, or $P(E \cup F)$ provided that the other three probabilities are known.

Example 6.22 Cable Services

Understand the context ➤ Suppose that 60% of all customers of a large cable company subscribe to Internet service, 40% subscribe to mobile phone service, and 25% have both types of services. If a customer is selected at random, what is the probability that they have at least one of these two types of service? Define the following events:

E = event that a selected customer has Internet service
F = event that a selected customer has mobile phone service

Do the work ➤ The given information implies that

$$P(E) = 0.60 \quad P(F) = 0.40 \quad P(E \cap F) = 0.25$$

from which you can calculate

$$\begin{aligned}&P(\text{customer has at least one of the two types of services})\\&\quad = P(E \cup F)\\&\quad = P(E) + P(F) - P(E \cap F)\\&\quad = 0.60 + 0.40 - 0.25\\&\quad = 0.75\end{aligned}$$

The event that the customer has neither type of service is $(E \cup F)^C$, so

$$P(\text{customer has neither type of service}) = 1 - P(E \cup F) = 0.25$$

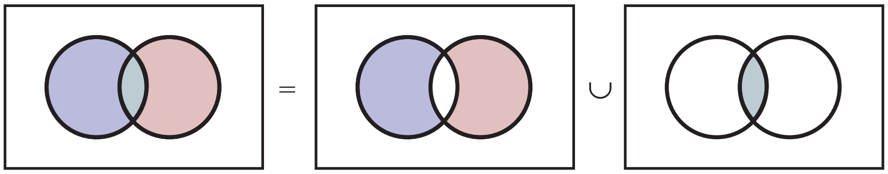
Now consider the probability that the selected customer has exactly one type of service. Referring to the Venn diagram in Figure 6.10, you can see that the event *at least one* can be thought of as consisting of two mutually exclusive parts: *exactly one* and *both*. This means that

$$\begin{aligned}P(E \cup F) &= P(\text{at least one})\\&= P(\text{exactly one} \cup \text{both})\\&= P(\text{exactly one}) + P(\text{both})\\&= P(\text{exactly one}) + P(E \cap F)\end{aligned}$$

It follows that

$$\begin{aligned}P(\text{exactly one}) &= P(E \cup F) - P(E \cap F)\\&= 0.75 - 0.25\\&= 0.50\end{aligned}$$

Figure 6.10
Representing $P(E \cup F)$ as the union of two mutually exclusive events.



The general addition rule for more than two events is rather complicated. For example, in the case of three events,

$$\begin{aligned}P(E \cup F \cup G) = P(E) + P(F) + P(G) - P(E \cap F) - P(E \cap G)\\ - P(F \cap G) + P(E \cap F \cap G)\end{aligned}$$

For more than three events, consult a text with more extensive coverage of probability.

Copyright 2025 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s). Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

General Multiplication Rule

In Section 6.4, the formula

$$P(E|F) = \frac{P(E \cap F)}{P(F)}$$

was used to calculate conditional probabilities when $P(E \cap F)$ is known. Sometimes, however, conditional probabilities are known or can be estimated. When this is the case, they can be used to calculate the probability of the intersection of two events.

Multiplying both sides of the conditional probability formula by $P(F)$ gives a useful expression for the probability that both events E and F will occur.

General Multiplication Rule for Two Events

For any two events E and F,

$$P(E \cap F) = P(E|F)P(F)$$

Example 6.23 Traffic School

Understand the context ➤

Suppose that 20% of all teenage drivers in a certain county received a citation for a moving violation in 2022. Also suppose that 80% of those receiving such a citation attended traffic school so that the citation would not appear on their driving record. If a teenage driver from this county is randomly selected, what is the probability that they received a citation and attended traffic school?

Formulate a plan ➤

Define two events E and F as follows:

E = selected driver attended traffic school
F = selected driver received a citation

The question posed can then be answered by calculating $P(E \cap F)$. The given percentages imply that

$$P(F) = 0.20$$
$$\text{and} \quad P(E|F) = 0.80.$$

Do the work ➤

Notice the difference between $P(E)$, which is the proportion in the entire population who attended traffic school (not given), and $P(E|F)$, which is the proportion of those receiving a citation who attended traffic school. Using the multiplication rule, you can calculate

$$\begin{aligned} P(E \text{ and } F) &= P(E|F)P(F) \\ &= (0.80)(0.20) \\ &= 0.16 \end{aligned}$$

Interpret the results ➤

This means that 16% of all teenage drivers in this county received a citation *and* attended traffic school.

Example 6.24 Smart Phone Warranties

Understand the context ➤

The following table gives information on smart phones sold by a large electronics store:

	Percentage of Customers Purchasing	**Of Those Who Purchase, Percentage Who Purchase Extended Warranty**
Brand 1	70	20
Brand 2	30	40

Copyright 2025 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s). Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

Suppose a purchaser is randomly selected from among all those who bought a smart phone from this store. What is the probability that the selected customer purchased a Brand 1 model and an extended warranty?

Formulate a plan ➤ To answer this question, first define the following events:

B_1 = event that Brand 1 is purchased
B_2 = event that Brand 2 is purchased
E = event that an extended warranty is purchased

The information in the table implies that

$$P(\text{Brand 1 purchased}) = P(B_1) = 0.70$$
$$P(\text{extended warranty}|\text{Brand 1 purchased}) = P(E|B_1) = 0.20$$

Do the work ➤ Notice that the 20% is identified with a *conditional* probability. *Among purchasers of Brand 1*, this is the percentage opting for an extended warranty. Substituting these numbers into the general multiplication rule, results in

$$\begin{aligned} P(B_1 \textit{ and } E) &= P(E|B_1)P(B_1) \\ &= (0.20)(0.70) \\ &= 0.14 \end{aligned}$$

The tree diagram of Figure 6.11 gives a nice visual display of how the general multiplication rule is used here. The two first-generation branches are labeled with events B_1 and B_2 along with their probabilities. Two second-generation branches extend from each first-generation branch. These correspond to the two events E and *not E*. The *conditional* probabilities $P(E|B_1)$, $P(\textit{not } E|B_1)$, $P(E|B_2)$, and $P(\textit{not } E|B_2)$, appear on these branches. Application of the multiplication rule then consists of multiplying probabilities along the branches of the tree diagram. For example,

$$\begin{aligned} P(\text{Brand 2 and warranty purchased}) &= P(B_2 \textit{ and } E) \\ &= P(B_2 \cap E) \\ &= P(E|B_2)P(B_2) \\ &= (0.4)(0.3) \\ &= 0.12 \end{aligned}$$

and this probability is displayed to the right of the E branch that comes from the B_2 branch.

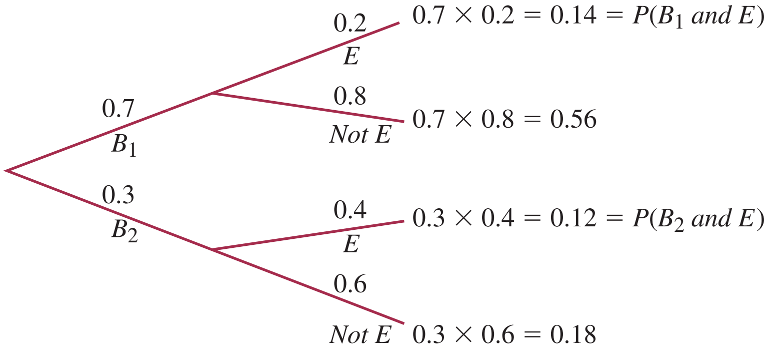


Figure 6.11
A tree diagram for the probability calculations of Example 6.24.

You can now easily calculate $P(E)$, the probability that an extended warranty is purchased. The event E can occur in two different ways: Buy Brand 1 *and* warranty, or buy Brand 2 *and* warranty. These events are $B_1 \cap E$ and $B_2 \cap E$. Furthermore, if each customer purchased a single smart phone, they could not have simultaneously purchased both Brand 1 and Brand 2, so the two events $B_1 \cap E$ and $B_2 \cap E$ are mutually exclusive.

Copyright 2025 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s). Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

It follows that

$$\begin{aligned} P(E) &= P(B_1 \cap E) + P(B_2 \cap E) \\ &= P(E|B_1)P(B_1) + P(E|B_2)P(B_2) \\ &= (0.2)(0.7) + (0.4)(0.3) \\ &= 0.14 + 0.12 \\ &= 0.26 \end{aligned}$$

Interpret the results ➤ This probability is the sum of two of the probabilities shown on the right-hand side of the tree diagram. This means that 26% of all smart phone purchasers selected an extended warranty.

The general multiplication rule can be extended to give an expression for the probability that several events occur together. In the case of three events E, F, and G,

$$P(E \cap F \cap G) = P(E|F \cap G)P(F|G)P(G)$$

When the events are all independent, $P(E|F \cap G) = P(E)$ and $P(F|G) = P(F)$ so the right-hand side of the equation for $P(E \cap F \cap G)$ simplifies to the product of the three unconditional probabilities.

Example 6.25 Lost Luggage

Understand the context ➤ Suppose that 20% of all passengers who fly from Los Angeles (LA) to New York (NY) do so on Airline G. This airline misplaces luggage for 10% of its passengers, and 90% of this lost luggage is found and returned to the owner. If a passenger who has flown from LA to NY is randomly selected, what is the probability that the selected individual flew on Airline G (event G), had luggage misplaced (event F), and the misplaced luggage was returned (event E)? The given information implies that

Do the work ➤

$$P(G) = 0.20 \quad P(F|G) = 0.10 \quad P(E|F \cap G) = 0.90$$

Then

$$P(E \cap F \cap G) = P(E|F \cap G)P(F|G)P(G) = (0.90)(0.10)(0.20) = 0.018$$

Interpret the results ➤ This means that about 1.8% of passengers flying from LA to NY fly on Airline G, have their luggage misplaced, and the lost luggage is returned.

Law of Total Probability

Let's reconsider the information on smart phone sales from Example 6.24. In this example, the following events were defined:

B_1 = event that Brand 1 is purchased
B_2 = event that Brand 2 is purchased
E = event that an extended warranty is purchased

Based on the information given in Example 6.24, the following probabilities are known:

$$P(B_1) = 0.7 \quad P(B_2) = 0.3 \quad P(E|B_1) = 0.2 \quad P(E|B_2) = 0.4$$

Notice that the conditional probabilities $P(E|B_1)$ and $P(E|B_2)$ are known but that the unconditional probability $P(E)$ is not known.

To find $P(E)$, notice that the event E can occur in two ways:

(1) A customer purchases an extended warranty *and* buys Brand 1 ($E \cap B_1$); or
(2) a customer purchases an extended warranty *and* buys Brand 2 ($E \cap B_2$).

Because these are the only ways in which E can occur, you can write the event E as

$$E = (E \cap B_1) \cup (E \cap B_2)$$

Copyright 2025 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s). Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

The two events $(E \cap B_1)$ and $(E \cap B_2)$ are mutually exclusive (since B_1 and B_2 are mutually exclusive), so using the addition rule for mutually exclusive events gives

$$\begin{aligned} P(E) &= P((E \cap B_1) \cup (E \cap B_2)) \\ &= P(E \cap B_1) + P(E \cap B_2) \end{aligned}$$

Finally, using the general multiplication rule to evaluate $P(E \cap B_1)$ and $P(E \cap B_2)$ results in

$$\begin{aligned} P(E) &= P(E \cap B_1) + P(E \cap B_2) \\ &= P(E|B_1)P(B_1) + P(E|B_2)P(B_2) \end{aligned}$$

Substituting in the known probabilities results in

$$\begin{aligned} P(E) &= P(E|B_1)P(B_1) + P(E|B_2)P(B_2) \\ &= (0.2)(0.7) + (0.4)(0.3) \\ &= 0.26 \end{aligned}$$

This means that 26% of the smart phone customers purchased an extended warranty.

This illustrates that when conditional probabilities are known, they can sometimes be used to calculate unconditional probabilities. The **law of total probability** formalizes this use of conditional probabilities.

The Law of Total Probability

If B_1 and B_2 are mutually exclusive events with $P(B_1) + P(B_2) = 1$, then for any event E

$$\begin{aligned} P(E) &= P(E \cap B_1) + P(E \cap B_2) \\ &= P(E|B_1)P(B_1) + P(E|B_2)P(B_2) \end{aligned}$$

More generally, if $B_1, B_2, \ldots, B_k$ are mutually exclusive events with $P(B_1) + P(B_2) + \cdots + P(B_k) = 1$, then for any event E

$$\begin{aligned} P(E) &= P(E \cap B_1) + P(E \cap B_2) + \cdots + P(E \cap B_k) \\ &= P(E|B_1)P(B_1) + P(E|B_2)P(B_2) + \cdots + P(E|B_k)P(B_k) \end{aligned}$$

Example 6.26 Which Way to Jump?

Understand the context ➤ The paper **"Action Bias among Elite Soccer Goalkeepers: The Case of Penalty Kicks" (*Journal of Economic Psychology* [2007]: 606–621)** presents an interesting analysis of 286 penalty kicks in televised championship soccer games from around the world. In a penalty kick, the only players involved are the kicker and the goalkeeper from the opposing team. The kicker tries to kick a ball into the goal from a point located 11 meters away. The goalkeeper tries to block the ball from reaching the goal.

For each penalty kick analyzed, the researchers recorded the direction that the goalkeeper moved (jumped to the left, stayed in the center, or jumped to the right) and whether or not the penalty kick was successfully blocked.

Consider the following events:

L = the event that the goalkeeper jumps to the left
C = the event that the goalkeeper stays in the center
R = the event that the goalkeeper jumps to the right
B = the event that the penalty kick is blocked

Consider the data ➤ Based on an analysis of the penalty kicks, the authors of the paper gave the following probability estimates:

$$P(B|L) = 0.142 \qquad P(B|C) = 0.333 \qquad P(B|R) = 0.126$$
$$P(L) = 0.493 \qquad P(C) = 0.063 \qquad P(R) = 0.444$$

Copyright 2025 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s). Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

Formulate a plan ➤

What proportion of penalty kicks were blocked? You can use the law of total probability to answer this question. Here, the three events L, C, and R play the role of B_1, B_2, and B_3 and B plays the role of E in the formula for the law of total probability.

Substituting into the formula, we get

Do the work ➤

$$\begin{aligned} P(B) &= P(B \cap L) + P(B \cap C) + P(B \cap R) \\ &= P(B|L)P(L) + P(B|C)P(C) + P(B|R)P(R) \\ &= (0.142)(0.493) + (0.333)(0.063) + (0.126)(0.444) \\ &= 0.070 + 0.021 + 0.056 \\ &= 0.147 \end{aligned}$$

Interpret the results ➤

This means that only 14.7% of penalty kicks were successfully blocked. Two other interesting findings of this study were

1. The direction that the goalkeeper moves appears to be independent of whether the kicker kicked the ball to the left, center, or right of the goal. This was attributed to the fact that goalkeepers have to choose their action before they can clearly observe the direction of the kick.
2. Based on the three conditional probabilities—$P(B|L) = 0.142$, $P(B|C) = 0.333$, and $P(B|R) = 0.126$—the optimal strategy for a goalkeeper appears to be to stay in the center of the goal. However, staying in the center was only chosen 6.3% of the time—much less often that jumping left or right. The authors believe that this is because a goalkeeper does not feel as bad about not successfully blocking a kick if some action (jumping left or right) is taken compared to if no action (staying in the center) is taken. This is the "action bias" referred to in the title of the paper.

Bayes' Rule

The discussion of probability rules is concluded by considering a formula discovered by the Reverend Thomas Bayes (1702–1761), an English Presbyterian minister. He discovered what is now known as Bayes' rule (or Bayes' theorem). Bayes' rule is a solution to what Bayes called the converse problem. To see what he meant by this, consider an example from the field of medical diagnosis.

Example 6.27 Lyme Disease

Understand the context ➤

Lyme disease is transmitted by infected ticks. Several tests are available for people with symptoms of Lyme disease. One of these tests is the EIA/IFA test. The paper **"Lyme Disease Testing by Large Commercial Laboratories in the United States" (*Clinical Infectious Disease* [2014]: 676–681)** found that 11.4% of those tested had Lyme disease.

Consider the following events:

$+$ represents a positive result on the test
$-$ represents a negative result on the test
L represents the event that the patient actually has Lyme disease
L^C represents the event that the patient actually does not have Lyme disease

For the population of people tested for Lyme disease, the following probabilities were reported in the article:

Consider the data ➤

Probability	**Interpretation**
$P(L) = 0.114$	The prevalence of Lyme disease in the population. About 11.4% of the population actually has Lyme disease.
$P(L^C) = 0.886$	88.6% of the population does not have Lyme disease.
$P(+|L) = 0.933$	93.3% of those with Lyme disease test positive.
$P(-|L) = 0.067$	6.7% of those with Lyme disease test negative.
$P(+|L^C) = 0.039$	3.9% of those who do not have Lyme disease test positive.
$P(-|L^C) = 0.961$	96.1% of those who do not have Lyme disease test negative.

Notice the form of the known conditional probabilities. For example, $P(+|L)$ is the probability of a positive test *given* that a person selected at random from the population actually has Lyme disease. Bayes' converse problem poses a question of a different form: Given that a person tests positive for the disease, what is the probability that they actually have Lyme disease? This converse problem is the one that is of primary interest in medical diagnosis problems.

Bayes reasoned as follows to obtain the answer to the converse problem of finding $P(L|+)$. From the definition of conditional probability, you know that

$$P(L|+) = \frac{P(L \cap +)}{P(+)}$$

Because $P(L \cap +) = P(+ \cap L)$, you can use the general multiplication rule to get

$$P(L \cap +) = P(+ \cap L) = P(+|L)P(L)$$

This helps, because both $P(+|L)$ and $P(L)$ are known. You now have

$$P(L|+) = \frac{P(+|L)P(L)}{P(+)}$$

The denominator $P(+)$ can be evaluated using the law of total probability, because L and L^C are mutually exclusive with $P(L) + P(L^C) = 1$. Applying the law of total probability to the denominator, you obtain

$$\begin{aligned} P(+) &= P(+ \cap L) + P(+ \cap L^C) \\ &= P(+|L)P(L) + P(+|L^C)P(L^C) \end{aligned}$$

You now have all you need to answer the converse problem:

Do the work ➤

$$\begin{aligned} P(L|+) &= \frac{P(+|L)P(L)}{P(+|L)P(L) + P(+|L^C)P(L^C)} \\ &= \frac{(0.933)(0.144)}{(0.933)(0.144) + (0.039)(0.886)} \\ &= \frac{0.106}{0.106 + 0.035} \\ &= 0.752 \end{aligned}$$

Interpret the results ➤

The probability $P(L|+)$ is a conditional probability. $P(L|+) = 0.752$ means that, in the long run, about 75.2% of those who test positive actually have the disease. Notice the difference between $P(L|+)$ and the previously reported conditional probability $P(+|L) = 0.933$, which means that 93.3% of those with Lyme disease test positive.

The accompanying box formalizes this reasoning in the statement of Bayes' rule.

Bayes' Rule

If B_1 and B_2 are mutually exclusive events with $P(B_1) + P(B_2) = 1$, then for any event E

$$P(B_1|E) = \frac{P(E|B_1)P(B_1)}{P(E|B_1)P(B_1) + P(E|B_2)P(B_2)}$$

More generally, if $B_1, B_2, \ldots, B_k$ are mutually exclusive events with $P(B_1) + P(B_2) + \cdots + P(B_k) = 1$ then for any event E,

$$P(B_i|E) = \frac{P(E|B_i)P(B_i)}{P(E|B_1)P(B_1) + P(E|B_2)P(B_2) + \cdots + P(E|B_k)P(B_k)}$$

Copyright 2025 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s). Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

Example 6.28 Contract Cheating

Understand the context ➤ Contract cheating is a form of plagiarism in which students pay someone else to complete an assignment for them. The authors of the paper **"How Prevalent is Contract Cheating and to What Extent are Students Repeat Offenders"** (***Journal of Academic Ethics*** **[2017]: 115–124)** collected data from college students by presenting them with the following scenario: "A student has to write an essay for their economics class. They find a website on the Internet where they can pay someone to write it for them. They pay the money and are emailed the essay. They hand the essay in as their own." Students are then asked if they had done a similar thing. Students were also classified according to whether they were a first-year college student.

Consider the data ➤ The following statements are based on the study results:

1. 46.5% of the study participants were first-year students
2. 2.1% of the study participants reported they had engaged in contract cheating
3. 36.4% of the students who reported they had engaged in contract cheating were first-year students.
4. 46.7% of the students who reported they had not engaged in contract cheating were first-year students.

Consider the chance experiment that consists of selecting one of the students participating in this survey at random, and define the following events:

F = the event that the selected student is a first-year student
NF = the event that the selected student is *not* a first-year student
C = the event that the selected student reports that they have engaged in contract cheating
NC = the event that the selected student reports that they have *not* engaged in contract cheating

The four statements based on the study results define the following probabilities:

$$P(F) = 0.465 \qquad P(NF) = 1 - 0.465 = 0.535$$
$$P(C) = 0.021 \qquad P(NC) = 1 - 0.021 = 0.979$$
$$P(F|C) = 0.364 \qquad P(F|NC) = 0.467$$

Formulate a plan ➤ Suppose that you want to know the proportion of those who reported contract cheating who were first-year students. This is equivalent to $P(C|F)$. Notice that you know $P(F|C)$ but not $P(C|F)$.

Do the work ➤ You can use Bayes' rule to evaluate $P(C|F)$ as follows (with C and NC playing the role of B_1 and B_2 and F playing the role of E in the formula for Bayes' rule):

$$\begin{aligned}
P(C|F) &= \frac{P(F|C)P(C)}{P(F|C)P(C) + P(F|NC)P(NC)} \\
&= \frac{(0.364)(0.021)}{(0.364)(0.021) + (0.467)(0.979)} \\
&= \frac{0.008}{0.008 + 0.457} \\
&= \frac{0.008}{0.465} \\
&= 0.017
\end{aligned}$$

Interpret the results ➤ This probability tells you that 1.7% of first-year students reported contract cheating.

What Do You Think?

What does the difference in the values of $P(C) = 0.021$ and $P(C|F) = 0.017$ tell you about contract cheating?

Exercises 6.69–6.87

6.69 Three of the most common types of pets are cats, dogs, and fish. Many families have more than one type of pet. Suppose that a family is selected at random and consider the following events and probabilities:

F = selected family has at least one fish
D = selected family has at least one dog
C = selected family has at least one cat

$$P(F) = 0.20 \qquad P(D) = 0.32 \qquad P(C) = 0.35$$
$$P(F \cap D) = 0.18 \quad P(F \cap C) = 0.07 \quad P(D \mid C) = 0.30$$

Calculate the following probabilities:

a. $P(F \mid D)$
b. $P(F \cup D)$
c. $P(C \cap D)$

6.70 A construction firm bids on two different contracts. Let E_1 be the event that the bid on the first contract is successful, and define E_2 to be the event that the bid on the second contract is successful. Suppose that $P(E_1) = 0.4$ and $P(E_2) = 0.3$ and that E_1 and E_2 are independent events.

a. Calculate the probability that both bids are successful (the probability of the event E_1 *and* E_2).
b. Calculate the probability that neither bid is successful (the probability of the event (*not* E_1) *and* (*not* E_2)).
c. What is the probability that the firm is successful in at least one of the two bids? (Hint: See Example 6.22.)

6.71 There are two traffic lights on Darlene's route from home to work. Let E denote the event that Darlene must stop at the first light, and define the event F in a similar manner for the second light. Suppose that $P(E) = 0.4$, $P(F) = 0.3$, and $P(E \cap F) = 0.15$.

a. What is the probability that Darlene must stop at at least one light; that is, what is the probability of the event $E \cup F$?
b. What is the probability that Darlene doesn't have to stop at either light?
c. What is the probability that Darlene must stop at exactly one of the two lights?
d. What is the probability that Darlene must stop just at the first light? (Hint: How is the probability of this event related to $P(E)$ and $P(E \cap F)$? A Venn diagram might help.)

6.72 Let F denote the event that a randomly selected registered voter in a certain city has signed a petition to recall the mayor. Also, let E denote the event that the randomly selected registered voter actually votes in the recall election. Describe the event $E \cap F$ in words. If $P(F) = 0.10$ and $P(E|F) = 0.80$, calculate $P(E \cap F)$. (Hint: See Example 6.23)

6.73 According to foodallergy.org, 8% of children in the United States have at least one food allergy. Seattle Children's Hospital reports that 40% of children with food allergies have a history of severe reactions.

a. If a child in the United States were to be randomly selected, what is the probability that they have at least one food allergy and have had a severe reaction?
b. Foodallergy.org also reported that 40% of children with food allergies are allergic to more than one food. If a child in the United States were to be randomly selected, what is the probability that they are allergic to multiple foods?

6.74 Suppose that Blue Cab operates 15% of the taxis in a certain city and Green Cab operates the other 85%. After a nighttime hit-and-run accident involving a taxi, an eyewitness said the vehicle was blue. Suppose, though, that under night vision conditions, only 80% of individuals can correctly distinguish between a blue and a green vehicle. What is the probability that the taxi at fault was blue? (Hint: A tree diagram might help.)

6.75 A large cable company reports the following:

- 80% of its customers subscribe to cable TV service
- 42% of its customers subscribe to Internet service
- 32% of its customers subscribe to telephone service
- 25% of its customers subscribe to both cable TV and Internet service
- 21% of its customers subscribe to both cable TV and phone service
- 23% of its customers subscribe to both Internet and phone service
- 15% of its customers subscribe to all three services

Consider the chance experiment that consists of selecting one of the cable company customers at random. Calculate and interpret the following probabilities:

a. P(cable TV only)
b. P(Internet | cable TV)
c. P(exactly two services)
d. P(Internet and cable TV only)

6.76 Refer to the information given in the previous exercise about customers of a large cable company.

a. Suppose two customers are to be selected at random. Would it be reasonable to consider the events C_1 = event that the first customer selected subscribes to cable TV and C_2 = event that the second customer selected subscribes to cable TV as independent events? Explain.

Copyright 2025 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s). Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

b. With C_1 and C_2 as defined in Part (a), calculate $P(C_1 \cap C_2)$.

6.77 The American Cancer Society reports that the lifetime risk of developing colon cancer is about 1 in 25. This means that about 4% of people develop colon cancer at some point during their lives. The good news is that this type of cancer is treatable if discovered early and there are several procedures that can be used to detect colon cancer and precancerous growths. The usual and most effective screening test is a colonoscopy, which uses a small camera at the end of a flexible tube to view the inside of the large intestine. In recent years, alternative screening procedures that are easier and not invasive have been developed. One of these, Cologuard, uses a stool sample that is collected at home and then sent to a lab for analysis. As with most lab tests, there is a possibility that the result is inaccurate. The maker of Cologuard has the following statement on its website **(cologuard.com, retrieved October 15, 2021):**

> False positives and false negative results can occur. In a clinical study, 13% of people without cancer received a positive result (false positive) and 8% of people with cancer received a negative result (false negative).

Consider the following events:

C = randomly selected person has colon cancer

not C = randomly selected person does not have colon cancer

$+$ = Cologuard test is positive

$-$ = Cologuard test is negative

Suppose that $P(C) = 0.04$.

a. What is the probability that a randomly selected person does not have colon cancer, $P(\text{not } C)$.

b. What is the probability that a randomly selected person tests positive given that they do not have colon cancer, $P(+|\text{not } C)$?

c. What is the probability that a randomly selected person tests negative given that they do not have colon cancer, $P(-|\text{not } C)$?

d. What is the probability that a randomly selected person tests negative given that they do have colon cancer, $P(-|C)$?

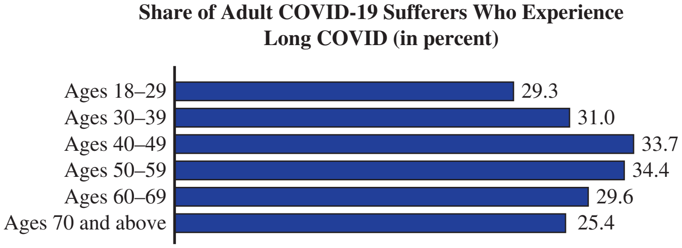
e. What is the probability that a randomly selected person tests positive given that they do have colon cancer, $P(+|C)$?

f. A probability that is of more interest for someone who tests positive is $P(C|+)$, the probability that a person actually has colon cancer given that they test positive. Calculate this probability.

g. Explain the difference between $P(+|C)$ and $P(C|+)$.

6.78 In May 2023, the U.S. Census Bureau **(census.gov)** published a report on the prevalence of long COVID (still experiencing symptoms more than 3 months after having had COVID) by age, educational attainment, and income. Overall, 31.1% of adult COVID-19 sufferers reported experiencing long COVID.

A figure similar to the one below appears on the census bureau website.



a. Each of the percentages in the given figure represents a conditional probability. Based on the figure, what is the probability of long COVID given that an adult with COVID-19 is in the 18 to 29 age group?

b. Are the two events C = event that an adult COVID-19 sufferer experienced long COVID and A = event that an adult COVID-19 sufferer is in the 18 to 29 age group independent?

c. According to statista.com, as of April 2023, the share of adult COVID-19 sufferers who are in the age group 18 to 29 was 24.4%. Calculate the probability that an adult COVID-19 sufferer is in the age group 18 to 29 and experienced long COVID.

d. Explain the difference between the probability calculated in Part (a) and the probability calculated in Part (c).

6.79 The report referenced in the previous exercise also indicated that there were observed differences in the probability of experiencing long COVID based on educational attainment. While overall 31.1 % of adult COVID-19 sufferers reported experiencing long COVID, the percentage was 40.9 for those who did not have a high school degree, 32.9 for those who had a high school degree but no college, 35.1 for those who had some college education but no college degree, and 23.6 for those who had a college degree.

a. What is the probability of long COVID given that an adult with COVID-19 did not complete high school?

b. What is the probability that an adult COVID-19 sufferer has a college degree and experienced long COVID?

c. Write a few sentences explaining how long COVID appears to be related to educational attainment.

Copyright 2025 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s). Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

6.80 The report **"Great Jobs, Great Lives. The Relationship Between Student Debt, Experiences and Perceptions of College Worth" (Gallup-Purdue Index 2015 Report)** gave information on the percentage of college graduates who graduated between 2006 and 2015 who strongly agreed with the statement "My college education was worth the cost." Suppose that person who graduated from college between 2006 and 2015 will be selected at random, and consider the following events:

S = event that the selected graduate strongly agrees that education was worth the cost
A_1 = event that the selected graduate had no student loan debt
A_2 = event that the selected graduate had student loan debt between \$1 and \$25,000
A_3 = event that the selected graduate had student loan debt between \$25,001 and \$50,000
A_4 = event that the selected graduate had student loan debt of more than \$50,000

The following probability estimates are from the report:

$P(A_1) = 0.37$ $P(A_2) = 0.28$ $P(A_3) = 0.22$ $P(A_4) = 0.13$
$P(S|A_1) = 0.49$ $P(S|A_2) = 0.43$ $P(S|A_3) = 0.30$
$P(S|A_4) = 0.18$

Use the Law of Total Probability to calculate $P(S)$.

6.81 Use Bayes' Rule and the information given in the previous exercise to answer the following questions.

a. Calculate $P(A_1|S)$.
b. Calculate $P(A_4|S)$.
c. Interpret the probabilities calculated in Parts (a) and (b).

6.82 The authors of the paper **"Do Physicians Know When Their Diagnoses Are Correct?" (*Journal of General Internal Medicine* [2005]: 334–339)** presented detailed case studies to medical students and to faculty at medical schools. Each participant was asked to provide a diagnosis in the case and also to indicate whether their confidence in the correctness of the diagnosis was high or low. Define the events C, I, and H as follows:

C = event that diagnosis is correct
I = event that diagnosis is incorrect
H = event that confidence in the correctness of the diagnosis is high

a. Data appearing in the paper were used to estimate the following probabilities for medical students:

$P(C) = 0.261$ $P(I) = 0.739$
$P(H|C) = 0.375$ $P(H|I) = 0.073$

Use Bayes' rule to calculate the probability of a correct diagnosis given that the student's confidence level in the correctness of the diagnosis is high.

b. Data from the paper were also used to estimate the following probabilities for medical school faculty:

$P(C) = 0.495$ $P(I) = 0.505$
$P(H|C) = 0.537$ $P(H|I) = 0.252$

Calculate $P(C|H)$ for medical school faculty. How does the value of this probability compare to the value of $P(C|H)$ for students computed in Part (a)?

6.83 Consider the following events:

C = event that a randomly selected driver is observed to be using a cell phone
A = event that a randomly selected driver is observed driving a passenger automobile
V = event that a randomly selected driver is observed driving a van or SUV
T = event that a randomly selected driver is observed driving a pickup truck

The probabilities given below are based on data from the **National Highway Traffic Safety Administration ("Traffic Safety Facts," February 2014):**

$P(C) = 0.015$ $P(C|A) = 0.017$
$P(C|V) = 0.014$ $P(C|T) = 0.010$

Explain why $P(C)$ is not just the average of the three given conditional probabilities.

6.84 The article **"U.S. Investors Split Between Digital and Traditional Banking" (gallup.com, August 5, 2016)** summarized data from a Gallup survey of a random sample of 1019 adult Americans with investments of \$10,000 or more. Based on the survey data, it was estimated that 31% of investors manage their investments by doing everything they possibly can online. But the authors of the article also noted that there was quite a difference between younger investors (age 18 to 49) and older investors (age 50 and older). For younger investors, 43% said they do everything they possibly can online, while the percentage for older investors was 23%.

a. Use the given information to estimate $P(O)$, $P(O|Y)$, and $P(O|F)$ where O = event that a randomly selected investor does everything possible online, Y = event that a randomly selected investor is age 18 to 49, and F = event that a randomly selected investor is 50 years old or older.

b. Suppose that 40% of investors are between the ages of 18 and 49. Use the probabilities from Part (a) and the estimate $P(Y) = 0.40$ to calculate $P(Y|O)$ and write a sentence interpreting this value.

Copyright 2025 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s). Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

6.85 Radiologists are often asked to predict the sex of a baby from ultrasound images made during pregnancy. The authors of the paper **"The Use of Three Dimensional Ultrasound for Fetal Gender Determination in the First Trimester" (*The British Journal of Radiology,* [2003]: 448–451)** followed up on 159 predictions made by a particular radiologist (Radiologist 1) to determine whether or not they were correct. Data from the paper is summarized in the accompanying table.

	Radiologist 1	
	Predicted Male	**Predicted Female**
Assigned Sex Is Male	74	12
Assigned Sex Is Female	14	59

a. Assuming that these data are representative of sex predictions made by Radiologist 1, estimate the probability that a sex prediction is correct, given that the baby is male.

b. Assuming that these data are representative of sex predictions made by Radiologist 1, estimate the probability that a sex prediction is correct, given that the baby is female.

c. For Radiologist 1, is a sex prediction more likely to be correct if the baby is male? Explain.

d. Estimate the probability that a sex prediction made by Radiologist 1 is correct.

6.86 The paper referenced in the previous exercise also included data for a second radiologist, Radiologist 2. Based on the data from the previous exercise for Radiologist 1 and the data in the accompanying table for Radiologist 2, write a paragraph comparing the accuracy of sex predictions made by these two radiologists.

	Radiologist 2	
	Predicted Male	**Predicted Female**
Assigned Sex Is Male	81	8
Assigned Sex Is Female	7	58

6.87 In an article that appears on the website of the American Statistical Association (**amstat.org**), Carlton Gunn, a public defender in Seattle, Washington, wrote about how they use statistics in their work as an attorney. They state:

> I personally have used statistics in trying to challenge the reliability of drug testing results. Suppose the chance of a mistake in the taking and processing of a urine sample for a drug test is just 1 in 100. And your client has a "dirty" (i.e., positive) test result. Only a 1 in 100 chance that it could be wrong? Not necessarily. If the vast majority of all tests given—say 99 in 100—are truly clean, then you get one false dirty and one true dirty in every 100 tests, so that half of the dirty tests are false.

Define the following events as:

TD = event that the test result is dirty
TC = event that the test result is clean
D = event that the sample tested is actually dirty
C = event that the sample tested is actually clean

a. Using the information in the quote, what are the values of

i. $P(TD|D)$ **iii.** $P(C)$
ii. $P(TD|C)$ **iv.** $P(D)$

b. Use the law of total probability to find $P(TD)$.

c. Use Bayes' rule to evaluate $P(C|TD)$. Is this value consistent with the argument given in the quote? Explain.

Section 6.7 Estimating Probabilities Empirically and Using Simulation

In the examples presented so far, reaching conclusions required knowledge of the probabilities of various outcomes. In some cases, this is reasonable, and you know the actual long-run proportion of the time that each outcome will occur. In other situations, these probabilities are not known. Sometimes probabilities can be determined analytically, by using the probability properties and probability rules introduced in this chapter. However, when this is not possible, impractical, or just beyond the limited probability tools of the introductory course, you can *estimate* probabilities empirically through observation or by simulation.

Estimating Probabilities Empirically

It is fairly common practice to use observed long-run proportions to estimate probabilities. The process of estimating probabilities empirically is simple:

Copyright 2025 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s). Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

1. Observe a very large number of chance outcomes under controlled circumstances.
2. Interpreting probability as a long-run relative frequency, estimate the probability of an event by using the observed proportion of occurrence.

This process is illustrated in Examples 6.29 and 6.30.

Example 6.29 Fair Hiring Practices

Understand the context ➤

The Biology Department at a university plans to recruit a new faculty member and intends to advertise for someone with a Ph.D. in biology and at least 10 years of college-level teaching experience. A member of the department expresses the belief that the experience requirement will exclude many potential applicants and will exclude far more applicants from underrepresented groups than from groups that are not underrepresented. The Biology Department would like to determine the probability that an applicant with a Ph.D. in biology would be eliminated from consideration because of the experience requirement.

A similar university completed a search in which there was no requirement for prior teaching experience, but the information about prior teaching experience was recorded. The 410 applications yielded the following data:

Consider the data ➤

	Number of Applicants		
Group	**Less Than 10 Years of Experience**	**10 Years of Experience or More**	**Total**
Not Underrepresented	178	112	**290**
Underrepresented	99	21	**120**
Total	**277**	**133**	**410**

Let's assume that the populations of applicants for the two positions can be regarded as the same. You can use the available information to approximate the probability that an applicant will fall into each of the four group–experience combinations.

Do the work ➤

The estimated probabilities (obtained by dividing the number of applicants for each group–experience combination by 410) are given in Table 6.1.

Table 6.1 Estimated Probabilities for Example 6.29

	Less Than 10 Years of Experience	**10 Years of Experience or More**
Not Underrepresented	0.434	0.273
Underrepresented	0.242	0.051

From Table 6.1, the estimate of P(applicant excluded because of the experience requirement) $= 0.434 + 0.242 = 0.676$.

Interpret the results ➤

You can also assess the impact of the experience requirement separately for applicants from groups that are not underrepresented and those that are from underrepresented groups. From the given information, the proportion of applicants from groups that are not underrepresented who have less than 10 years of experience is $178/290 = 0.614$, whereas the corresponding proportion for applicants from underrepresented groups is $99/120 = 0.825$. This means that approximately 61% of the applicants who are not from underrepresented groups would be eliminated by the experience requirement, and about 83% of the applicants from underrepresented groups would be eliminated.

These subgroup proportions—0.614 for applicants who are not from underrepresented groups and 0.825 for applicants from underrepresented groups—are estimates of conditional probabilities, which show how the original probability changes in light of new information. In this example, the probability that a potential applicant has less than 10 years of experience is 0.676, but this probability changes to 0.825 if we know that a applicant is from an underrepresented group. These probabilities can be expressed as

Copyright 2025 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s). Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

$P(\text{less than 10 years of experience}) = 0.676$ (an unconditional probability)

and

$P(\text{less than 10 years of experience} \mid \text{applicant is from an underrepresented group}) = 0.825$ (a conditional probability)

Example 6.30 An Unexpected Result

Understand the context ➤

The article **"People With Covid Jabs Have Been Less Likely to Die of Other Causes" (*The Economist*, October 27, 2021)** summarizes data from a study carried out by the Centers for Disease Control and Prevention. In this study, a large number of people were followed for a year. The study included both people who were fully vaccinated for COVID-19 and people who were not vaccinated. For any people in the study who died, the cause of death was recorded. Any death that was attributed to COVID-19 was excluded for both vaccinated and unvaccinated people.

Consider the data ➤

Data consistent with the mortality rates from the study for people vaccinated with the Pfizer vaccine and for unvaccinated people are given Table 6.2 (the sample sizes in the actual study were much larger).

Table 6.2 Non-COVID Related Deaths for Pfizer Vaccine vs. No Vaccination

	Fully Vaccinated	Not Vaccinated	Total
Died, Not COVID Related	7	23	**30**
Did Not Die During Study Period	1993	1977	**3970**
Total	**2000**	**2000**	**4000**

Do the work ➤

Because the study participants were selected to be representative of American adults, you can use the information provided by the table to estimate various probabilities. For example, if an adult is selected at random, the probability that they might die during the study period due to a non-COVID-related cause is

$$P(\text{non-COVID related death}) = \frac{30}{4000} = 0.008$$

However, the probability of a non-COVID-related death given that the selected person was fully-vaccinated with the Pfizer vaccine is

$$P(\text{non-COVID related death} \mid \text{vaccinated}) = \frac{7}{2000} = 0.0035$$

Also,

$$P(\text{non-COVID related death} \mid \text{not vaccinated}) = \frac{23}{2000} = 0.0155$$

which is more than three times the probability for the vaccinated. Similar probabilities were reported for the Moderna vaccine.

Interpret the results ➤

While it was expected that deaths due to COVID-19 would be lower in the vaccinated group, it was surprising to see lower mortality rates due to non-COVID deaths as well. The article correctly cautions again drawing a cause-and-effect conclusion because this was an observational study where participants were not randomly assigned to get vaccinated or to not get vaccinated. They point out that there could be other factors that could contribute to the observed difference in mortality rates.

What Do You Think?

What other variables might be considered as potentially confounding variables that would make you cautious about drawing a cause-and-effect conclusion?

The following quote from the paper is informative and addresses this issue:

> Of course, there are other reasons why the health of the two groups may differ systematically. Some people are advised not to get jabs because of pre-existing medical conditions, so that the unvaccinated cohort is likely to contain more people with complex medical histories. During a pandemic, the unvaccinated may also have extra reason to delay seeking all kinds of medical care, for fear of catching the virus in doctors' offices or hospitals. This could result in deadly conditions like cancer or heart disease being detected only too late.
>
> Large observational studies can give unparalleled insights into the safety and efficacy of vaccines across diverse populations. However, real-world data is never perfectly controlled. It seems all but certain that some still-invisible difference between people who get the vaccine and those who do not, rather than some unknown benefit of the jab, is to thank (or blame) for the vaccine's correlative effects. Jabs are not immunizations against mortality.

Estimating Probabilities Using Simulation

Simulation provides a way to estimate probabilities when you are unable to determine probabilities analytically and when it is impractical to estimate them empirically by observation. Simulation is a method that generates "observations" by performing a chance experiment that is as similar as possible in structure to the real situation of interest.

To illustrate the idea of simulation, consider the situation in which a professor wishes to estimate the probabilities of different possible scores on a 20-question true–false quiz when students are just guessing at the answers. Because each question is a true–false question, a person who is guessing should be equally likely to answer correctly or incorrectly on any given question.

Rather than asking a guessing student to select true or false and then comparing the choice to the correct answer, an equivalent process would be to pick a ball at random from a box that contains half red balls and half blue balls, with a blue ball representing a correct answer. Making 20 selections from the box (with replacement) and then counting the number of correct choices (the number of times a blue ball is selected) is a physical substitute for an observation from a student who has guessed at the answers to 20 true–false questions. Any particular number of blue balls in 20 selections should have the same probability as the same number of correct responses to the quiz when a student is guessing.

For example, 20 selections of balls might produce the following results:

Selection	1	2	3	4	5	6	7	8	9	10
	R	R	B	R	B	B	R	R	R	B

Selection	11	12	13	14	15	16	17	18	19	20
	R	R	B	R	R	B	B	R	R	B

Because 8 of the 20 selections resulted in a blue ball, this would correspond to a quiz with eight correct responses, and it would provide you with one observation for estimating the probabilities of interest. This process could then be repeated a large number of times to generate additional observations. For example, you might find the following:

Repetition	Number of "Correct" Responses
1	8
2	11
3	10
4	12
⋮	⋮
1000	11

The 1000 simulated quiz scores could then be used to construct a table of estimated probabilities.

Copyright 2025 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s). Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

Taking this many balls out of a box and writing down the results would be cumbersome and tedious. The process can be simplified by using random digits to substitute for drawing balls from the box. For example, a single digit could be selected at random from the 10 digits 0, 1, 2, 3, 4, 5, 6, 7, 8, 9. When using random digits, each of the 10 possibilities is equally likely to occur, so you can use the even digits (including 0) to indicate a correct response and the odd digits to indicate an incorrect response. This would maintain the important property that a correct response and an incorrect response are equally likely, because correct and incorrect are each represented by 5 of the 10 digits.

To aid in carrying out such a simulation, tables of random digits (such as Appendix A Table 1) or computer-generated random digits can be used. The numbers in Appendix A Table 1 were generated using a computer's random number generator.

To see how a table of random digits can be used to carry out a simulation, reconsider the quiz example. You can use a random digit to represent the guess on a single question, with an even digit representing a correct response. A sequence of 20 digits could represent the answers to the 20 quiz questions. Start by picking an arbitrary starting point in Appendix A Table 1. Suppose that you start at row 10 and take the 20 digits in a row to represent one quiz. The first five "quizzes" and the corresponding number correct (number of even digits) are:

Quiz	Random Digits	Number Correct
1	9 4 6 0 6 9 7 8 8 2 5 2 9 6 0 1 4 6 0 5	**13**
2	6 6 9 5 7 4 4 6 3 2 0 6 0 8 9 1 3 6 1 8	**12**
3	0 7 1 7 7 7 2 9 7 8 7 5 8 8 6 9 8 4 1 0	**9**
4	6 1 3 0 9 7 3 3 6 6 0 4 1 8 3 2 6 7 6 8	**11**
5	2 2 3 6 2 1 3 0 2 2 6 6 9 7 0 2 1 2 5 8	**13**

This process would be repeated to generate a large number of observations, which would then be used to construct a table of estimated probabilities.

The method for generating observations must preserve the important characteristics of the actual process being considered if simulation is to be successful. For example, it would be easy to adapt the simulation procedure for the true–false quiz to one for a multiple-choice quiz. Suppose that each of the 20 questions on the quiz has five possible responses, only one of which is correct. For any particular question, you would expect a student to be able to guess the correct answer only one-fifth of the time in the long run.

To simulate this situation, you could select at random from a box that contained four red balls and only one blue ball (or, more generally, four times as many red balls as blue balls). If you are using random digits for the simulation, you could use 0 and 1 to represent a correct response and 2, 3, . . . , 9 to represent an incorrect response.

Using Simulation to Approximate a Probability

1. Propose a method that uses a random mechanism (such as a random number generator or table, the selection of a ball from a box, the toss of a coin, etc.) to represent an observation. Be sure that the important characteristics of the actual process are preserved.
2. Generate an observation using the method from Step 1, and determine whether the outcome of interest has occurred.
3. Repeat Step 2 a large number of times.
4. Calculate the estimated probability by dividing the number of observations for which the outcome of interest occurred by the total number of observations generated.

The simulation process is illustrated in Examples 6.31–6.33.

Example 6.31 Building Permits

Understand the context ➤

Many cities in California limit the number of building permits that are issued each year. Because of limited water resources, one such city plans to issue permits for only 10 dwelling units in the upcoming year. The city will decide who is to receive permits by holding a lottery.

Suppose that you are one of 39 individuals who apply for permits. Thirty of these individuals are requesting permits for a single-family home, eight are requesting permits for a duplex (which counts as two dwelling units), and one person is requesting a permit for a small apartment building with eight units (which counts as eight dwelling units). Each request will be entered into the lottery.

Requests will be selected at random one at a time, and if there are enough permits remaining, the request will be granted. This process will continue until all 10 permits have been issued. If your request is for a single-family home, what is the probability that you receive a permit? You can use simulation to estimate this probability. (It is not easy to determine analytically.)

To carry out the simulation, you can view the requests as being numbered from 1 to 39 as follows:

01–30	Requests for single-family homes
31–38	Requests for duplexes
39	Request for 8-unit apartment

For ease of discussion, assume that your request is number 01.

One method for simulating the permit lottery consists of these three steps:

Formulate a plan ➤

1. Choose a random number between 01 and 39 to indicate which permit request is selected first, and grant this request.
2. Select another random number between 01 and 39 to indicate which permit request is considered next. Determine the number of dwelling units for the selected request. Grant the request only if there are enough permits remaining to satisfy the request.
3. Repeat Step 2 until permits for 10 dwelling units have been granted.

Do the work ➤

Minitab was used to generate random numbers between 01 and 39 to imitate the lottery drawing. (The random number table in Appendix A Table 1 could also be used by selecting two digits and ignoring 00 and any value over 39.) The first sequence generated by Minitab was

Random Number	Type of Request	Total Number of Units So Far
25	Single-family home	1
07	Single-family home	2
38	Duplex	4
31	Duplex	6
26	Single-family home	7
12	Single-family home	8
33	Duplex	10

You would stop at this point, because permits for 10 units would have been issued. In this simulated lottery, Request 01 was not selected, so you would not have received a permit.

The next simulated lottery (using Minitab to generate the selections) was as follows:

Random Number	Type of Request	Total Number of Units So Far
38	Duplex	2
16	Single-family home	3
30	Single-family home	4
39	Apartment—not granted, since there are not 8 permits remaining	4
14	Single-family home	5
26	Single-family home	6
36	Duplex	8
13	Single-family home	9
15	Single-family home	10

Copyright 2025 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s). Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

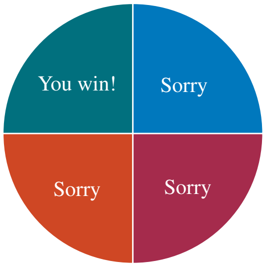
Again, Request 01 was not selected, so you would not have received a permit in this simulated lottery.

Now that a strategy for simulating a lottery has been devised, the tedious part of the simulation begins. You would now simulate a large number of lottery drawings, determining for each one whether Request 01 was granted. After 500 simulated lotteries were conducted, Request 01 was selected in 85 of the lotteries. This results in

$$\text{estimated probability of receiving a building permit} = \frac{85}{500} = 0.17$$

Example 6.32 Carnival Game Play

Understand the context ➤ Suppose that a carnival game allows you to buy a chance to spin a wheel like the one shown below that is divided into four equal sections. Three of the sections are labeled “Sorry” and the other one is labeled “You Win!”



If your spin results in “You Win!”, you win a prize valued at $5. You really want the prize and you plan to play the game until you win it. You can use simulation to estimate various probabilities. For example, suppose that it costs $2 for each spin of the wheel. Since the prize is valued at $5, you are hoping that you will win the prize on your first or second spin. What is the probability that this happens? Or you might wonder what is the probability that you will need 5 or more spins to win the prize?

Formulate a plan ➤ To simulate playing the game until you win a prize, you can use a single-digit random number to represent a spin of the wheel. A statistics software package can be used to generate random digits between 1 and 4. Since the probability of winning on any one spin is $\frac{1}{4}$, you can use 1, 2, and 3 to represent the wheel landing on a “Sorry” segment and 4 to represent the wheel landing on “You Win!” An observation is constructed by selecting a sequence of random digits between 1 and 4. If the first selected number is 4, then you win in one spin and the observation is complete. If the first selected digit is a 1, 2, or 3, you lost the first spin and another digit would be selected to represent the second spin. You would continue this way until a 4 is obtained.

Do the work ➤ Ten random digits generated using technology are

$$\underline{4}\qquad \underline{1\qquad 1\qquad 1\qquad 4}\qquad \underline{3\qquad 4}\qquad \underline{4}\qquad 2\qquad 2$$

Using these numbers to simulate observations results in the following:

Observation 1:	4	win in 1 spin
Observation 2:	1 1 1 4	win in 4 spins
Observation 3:	3 4	win in 2 spins
Observation 4:	4	win in 1 spin

After simulating four observations, there were 3 where a win occurred in 2 or fewer spins. Continuing this simulation to obtain 100 observations resulted in the observations summarized in Table 6.3.

Copyright 2025 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s). Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

Table 6.3 Simulated Outcomes for Carnival Game of Example 6.32

Number of Spins	Number of Outcomes
1	24
2	24
3	11
4	11
5	8
6	2
7	4
8	3
9	5
10	2
11	3
12	0
13	1
14	1
15	1
Total	**100**

Interpret the results ➤ Out of the 100 simulated observations, 48 resulted in a win in 2 or fewer spins, so the estimated probability would be 0.48. If you play this game until you win the prize, you pay more than the $5 that the prize is worth about half the time, and sometimes you would pay much more than the prize is worth. Forty-one of the simulated outcomes resulting in 5 or more spins needed to win the prize, so the estimated probability that it would take 5 or more spins to win is 0.41.

What Do You Think?

Would you play this game, spinning until you won the prize? Would you play if the cost per spin were $1 rather than $2?

Example 6.33 ESP?

Understand the context ➤ Can a close friend read your mind? Try the following chance experiment. Write the word *blue* on one piece of paper and the word *red* on another, and place the two slips of paper in a box. Select one slip of paper from the box, look at the word written on it, and then try to convey the word by sending a mental message to a friend who is seated in the same room. Ask your friend to select either red or blue, and record whether the response is correct. Repeat this 9 more times and determine the number of correct responses. How did your friend do? Is your friend receiving your mental messages or just guessing?

Formulate a plan ➤ You can investigate by using simulation to get the approximate probabilities of the various possible numbers of correct responses for someone who is guessing. Someone who is guessing should have an equal chance of responding correctly or incorrectly. You can use a random digit to represent a response, with an even digit representing a correct response (C) and an odd digit representing an incorrect response (X). A sequence of 10 digits can be used to simulate performing the chance experiment one time.

Do the work ➤ For example, using the last 10 digits in row 25 of the random number table (Appendix A Table 1) gives

5	2	8	3	4	3	0	7	3	5
X	C	C	X	C	X	C	X	X	X

which is a simulated chance experiment resulting in four correct responses. Minitab was used to generate 150 sequences of 10 random digits, resulting in the following results:

Copyright 2025 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s). Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

Sequence Number	Digits	Number Correct
1	3996285890	**5**
2	1690555784	**3**
3	9133190550	**2**
⋮	⋮	⋮
149	3083994450	**5**
150	9202078546	**7**

Table 6.4 summarizes the results of the simulation.

Interpret the results ➤

The estimated probabilities in Table 6.4 are based on the assumption that a correct and an incorrect response are equally likely (guessing). Evaluate your friend's performance in light of the information in Table 6.4. Is it likely that someone who is guessing would have been able to get as many correct as your friend did? Do you think your friend was receiving your mental messages? How are the estimated probabilities in Table 6.4 used to support your answer?

Table 6.4 **Estimated Probabilities for Example 6.33**

Number Correct	Number of Sequences	Estimated Probability
0	0	0.000
1	1	0.007
2	8	0.053
3	16	0.107
4	30	0.200
5	36	0.240
6	35	0.233
7	17	0.113
8	7	0.047
9	0	0.000
10	0	0.000
Total	**150**	**1.000**

Exercises 6.88–6.99

6.88 The report **"September 2021 Air Travel Consumer Report"** provides an overview of the complaints about airlines received by the U.S. Department of Transportation. The accompanying table gives the number of complaints received by type of complaint for July 2020 and July 2021.

Type of Complaint	July 2020 Number of Complaints Received	July 2021 Number of Complaints Received
Flight problems	84	875
Baggage handling	48	227
Reservations, ticketing, boarding	163	533
Customer service	112	203
Fares	367	462
Refunds	10,261	2,543

Suppose that one of these complaints is randomly selected for a follow-up interview. Use the given information to estimate the following probabilities.

a. The probability that a complaint made in July 2020 was about baggage handling.

b. The probability that a complaint made in July 2021 was not about flight problems.

c. The probability that two independent complaints made in July 2021 were both about flight problems.

d. The probability that a complaint made in July 2020 was either about flight problems or customer service.

6.89 Five hundred first-year students at a state university were classified according to both high school GPA and whether they were on academic probation at the end of their first semester. The data are summarized in the accompanying table.

	High School GPA			
Probation	**2.5 to <3.0**	**3.0 to <3.5**	**3.5 and Above**	**Total**
Yes	50	55	30	**135**
No	45	135	185	**365**
Total	**95**	**190**	**215**	**500**

a. Construct a table of the estimated probabilities for each GPA–probation combination. (Hint: See Example 6.25.)

b. Use the table constructed in Part (a) to approximate the probability that a randomly selected first-year student at this university will be on academic probation at the end of the first semester.

c. What is the estimated probability that a randomly selected first-year student at this university had a high school GPA of 3.5 or above?

d. Are the two events *selected student has a high school GPA of 3.5 or above* and *selected student is on academic probation at the end of the first semester* independent events? How can you tell?

6.90 Use the information given in the previous exercise to answer the following questions.

a. Estimate the proportion of first-year students with high school GPAs between 2.5 and 3.0 who are on academic probation at the end of the first semester.

b. Estimate the proportion of those first-year students with high school GPAs 3.5 and above who are on academic probation at the end of the first semester.

6.91 The report **"Great Jobs, Great Lives. The Relationship Between Student Debt, Experiences and Perceptions of College Worth" (Gallup-Purdue Index 2015 Report)** summarized data from a survey of 30,151 college graduates who graduated between 2006 and 2015. Each student surveyed responded to the statement "My college education was worth the cost." Students were also asked how much student loan debt they had taken on by the time they graduated. The student responses are summarized in the table at the bottom of the page.

Suppose that the data in this table is representative of college graduates.

a. What is the probability that a randomly selected college graduate had no student loan debt?

b. What is the probability that a randomly selected college graduate either disagrees or strongly disagrees that their education was worth the cost?

c. What is the probability that a randomly selected college graduate agrees or strongly agrees that their education was worth the cost and had no student loan debt?

d. What is the probability that a randomly selected college graduate agrees that their education was worth the cost, given that they had no student loan debt?

6.92 The U.S. Census Bureau issued a press release in 2022 titled **"School Enrollment Rises Slightly in 2021" (census.gov)**. Data on first-year full-time college students in Fall of 2021 from this report was used to construct the accompanying table. Age and type of institution was recorded for each first-year student. The table gives the number of students falling into each age group/institution type category.

Age Group	**Enrolled at a Two-Year College**	**Enrolled at a Four-Year College**
15–19	681,000	1,451,000
20–24	182,000	338,000
25–34	120,000	111,000
35 and older	52,000	61,000

a. If a first-year student is selected at random from those attending four-year colleges, what is the probability that the selected student is in the 20–24 age group?

b. If a first-year student is selected at random, what is the probability that the selected student is in the 20–24 age group?

Table for Exercise 6.91

	Response					
Debt	**Strongly Disagree**	**Disagree**	**Neither Agree nor Disagree**	**Agree**	**Strongly Agree**	**Row Total**
None	446	558	1,339	3,235	5,578	11,156
$1 to $25,000	422	844	1,435	2,111	3,630	8,442
$25,001 to $50,000	531	796	1,459	1,857	1,990	6,633
More than $50,000	706	745	862	901	706	3,920
Column Total	2,105	2,943	5,095	8,104	11,904	30,151

c. If a first-year student is selected at random, what is the probability that the selected student is enrolled at a four-year college, given that they are in the 35 and older age group?
d. If a first-year student is selected at random, what is the probability that the selected student is in the 20–24 age group and is enrolled at a two-year college?

6.93 Refer to the information given in the previous exercise.
a. If a first-year student is selected at random, what is the probability that the selected student is either in the 15–19 age group or is 35 or older?
b. If a first-year student is selected at random, what is the probability that the selected student is enrolled at a two-year college?
c. If two first-year students are selected at random, what is the probability that both are enrolled at four-year colleges?
d. If two first-year students are selected at random, what is the probability that neither is enrolled at a four-year college?

6.94 Refer to the information given in Exercises 6.92 and 6.93.
a. If two first-year students are selected at random, what is the probability that exactly one is in the 25–34 age group?
b. If two first-year students are selected at random, what is the probability that both are in the same age group?
c. If two first-year students are selected at random, what is the probability that they are from different age groups?

6.95 A medical research team wishes to evaluate two different treatments for a disease. Subjects are selected two at a time, and then one of the two is assigned to Treatment 1 and the other is assigned to Treatment 2. The treatments are applied, and each is either a success (S) or a failure (F).

The researchers keep track of the total number of successes for each treatment. They plan to continue the chance experiment until the number of successes for one treatment exceeds the number of successes or the other treatment by 2. For example, they might observe the results in the table at the bottom of the page. The chance experiment would stop after the sixth pair, because Treatment 1 has two more successes than Treatment 2. The researchers would conclude that Treatment 1 is preferable to Treatment 2.

Suppose that Treatment 1 has a success rate of 0.7 (that is, $P(\text{success}) = 0.7$ for Treatment 1) and that Treatment 2 has a success rate of 0.4. Use simulation to estimate the probabilities in Parts (a) and (b) by using the following procedure:

1. Use a pair of random digits to simulate one pair of subjects. Let the first digit represent Treatment 1 and use 1–7 as an indication of a success and 8, 9, and 0 to indicate a failure. Let the second digit represent Treatment 2, with 1–4 representing a success. For example, if the two digits selected to represent a pair were 8 and 3, you would record failure for Treatment 1 and success for Treatment 2.
2. Continue to select pairs, keeping track of the total number of successes for each treatment. Stop the trial as soon as the number of successes for one treatment exceeds that for the other by 2. This would complete one trial.
3. Repeat this whole process until you have results for at least 20 trials (more is better).
4. Use the simulation results to estimate the desired probabilities.
   a. Estimate the probability that more than five pairs must be treated before a conclusion can be reached. (Hint: $P(\text{more than } 5) = 1 - P(5 \text{ or fewer})$.)
   b. Estimate the probability that the researchers will incorrectly conclude that Treatment 2 is the better treatment.

6.96 Many cities regulate the number of taxi licenses, and there is a great deal of competition for both new and

Table for Exercise 6.95

Pair	Treatment 1	Treatment 2	Total Number of Successes for Treatment 1	Total Number of Successes for Treatment 2
1	S	F	1	0
2	S	S	2	1
3	F	F	2	1
4	S	S	3	2
5	F	F	3	2
6	S	F	4	2

Copyright 2025 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s). Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

existing licenses. Suppose that a city has decided to sell 10 new licenses for $25,000 each. A lottery will be held to determine who gets the licenses, and no one may request more than three licenses.

Twenty individuals and taxi companies have entered the lottery. Six of the 20 entries are requests for 3 licenses, 9 are requests for 2 licenses, and the rest are requests for a single license. The city will select requests at random, filling as many of the requests as possible.

For example, the city might fill requests for 2, 3, 1, and 3 licenses and then select a request for 3. Because there is only one license left, the last request selected would receive a license, but only one.

a. An individual has put in a request for a single license. Use simulation to approximate the probability that the request will be granted. Perform *at least* 20 simulated lotteries (more is better!).

b. Do you think that this is a fair way of distributing licenses? Can you propose an alternative procedure for distribution?

6.97 Four students must work together on a group project. They decide that each will take responsibility for a particular part of the project, as follows:

Person	Maria	Alex	Juan	Jacob
Task	Survey design	Data collection	Analysis	Report writing

Because of the way the tasks have been divided, one student must finish before the next student can begin work.

To ensure that the project is completed on time, a schedule is established, with a deadline for each team member. If any one of the team members is late, the timely completion of the project is jeopardized. Assume the following probabilities:

1. The probability that Maria completes on time is 0.8.
2. If Maria completes on time, the probability that Alex completes on time is 0.9, but if Maria is late, the probability that Alex completes on time is only 0.6.
3. If Alex completes on time, the probability that Juan completes on time is 0.8, but if Alex is late, the probability that Juan completes on time is only 0.5.
4. If Juan completes on time, the probability that Jacob completes on time is 0.9, but if Juan is late, the probability that Jacob completes on time is only 0.7.

Use simulation (with at least 20 trials) to estimate the probability that the project will be completed on time. Think carefully about this one. For example, you might use a random digit to represent each part of the project (four in all). For the first digit (Maria's part), 1–8 could represent *on time* and 9 and 0 could represent *late*. Depending on what happened with Maria (late or on time), you would then look at the digit representing Alex's part. If Maria was on time, 1–9 would represent *on time* for Alex, but if Maria was late, only 1–6 would represent *on time*. The parts for Juan and Jacob could be handled similarly.

6.98 In Exercise 6.97, the probability that Maria completes on time was 0.8. Suppose that this probability is really only 0.6. Use simulation (with at least 20 trials) to estimate the probability that the project will be completed on time.

6.99 Refer to Exercises 6.97 and 6.98. Suppose that the probabilities of timely completion are as in Exercise 6.97 for Maria, Alex, and Juan, but that Jacob has a probability of completing on time of 0.7 if Juan is on time and 0.5 if Juan is late.

a. Use simulation (with at least 20 trials) to estimate the probability that the project will be completed on time.

b. Compare the probability from Part (a) to the one computed in Exercise 6.98. Which decrease in the probability of on-time completion (Maria's or Jacob's) made the bigger change in the probability that the project will be completed on time?

Chapter Activities

Activity 6.1 Kisses

The purpose of this activity is to work as a class to develop and implement a plan to estimate a probability.

Background: The paper **"What Is the Probability of a Kiss? (It's Not What You Think)" (*Journal of Statistics Education* (online) [2002])** posed the following question: What is the probability that a Hershey's Kiss will land on its base (as opposed to its side) if it is flipped onto a table? Unlike flipping a coin, there is no reason to believe that this probability would be 0.5.

Working as a class, develop a plan that would enable you to estimate this probability empirically.

Once you have an acceptable plan, carry it out and use the resulting data to produce an estimate of the desired probability. Do you think that a kiss is equally likely to land on its base or on its side? Explain.

Copyright 2025 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s). Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

Activity 6.2 A Crisis for European Sports Fans?

The purpose of this activity is to use simulation to estimate a probability that can be used to investigate a claim.

Background: The *New Scientist* (January 4, 2002) reported on a controversy surrounding the Euro coins that have been introduced as a common currency across Europe. Each country mints its own coins, but these coins are accepted in any of the countries that have adopted the Euro as their currency.

A group in Poland claims that the Belgium-minted Euro does not have an equal chance of landing heads or tails. This claim was based on 250 tosses of the Belgium-minted Euro, of which 140 (56%) came up heads. Should this be cause for alarm for European sports fans, who know that "important" decisions are made by the flip of a coin?

In this activity, you will investigate whether this should be cause for alarm by examining whether observing 140 heads out of 250 tosses is an unusual outcome if the coin is fair.

1. For this first step, you can either (a) flip a U.S. penny 250 times, keeping a tally of the number of heads and tails observed (this won't take as long as you think), or (b) simulate 250 coin tosses by using your calculator or a statistics software package to generate random numbers (if you choose this option, give a brief description of how you carried out the simulation).
2. For your sequence of 250 tosses, calculate the proportion of heads observed.
3. Form a data set that consists of the values for proportion of heads observed in 250 tosses of a fair coin for the entire class. Summarize this data set by constructing a graphical display.
4. Working with a partner, write a paragraph explaining why European sports fans should or should not be worried by the results of the Polish experiment. Your explanation should be based on the observed proportion of heads from the Polish experiment and the graphical display constructed in Step 3.

Activity 6.3 The "Hot Hand" in Basketball

The purpose of this activity is to use simulation to estimate a probability.

Background: Consider a mediocre basketball player who has consistently made only 50% of their free throws over several seasons. If you were to examine their free throw record over the last 50 free throw attempts, is it likely that you would see a streak of 5 in a row where the player is successful in making the free throw? In this activity, you will investigate this question. Assume that the outcomes of successive free throw attempts are independent and that the probability that the player is successful on any particular attempt is 0.5.

1. Begin by simulating a sequence of 50 free throws for this player. Because this player has probability of success of 0.5 for each attempt and the attempts are independent, you can model a free throw by tossing a coin. Using heads to represent a successful free throw and tails to represent a missed free throw, simulate 50 free throws by tossing a coin 50 times, recording the outcome of each toss.
2. For your sequence of 50 tosses, identify the longest streak by looking for the longest string of heads in your sequence. Determine the length of this longest streak.
3. Combine your longest streak value with those from the rest of the class and construct a histogram or dotplot of these longest streak values.
4. Based on the graph from Step 3, does it appear likely that a player of this skill level would have a streak of 5 or more successes sometime during a sequence of 50 free throw attempts? Justify your answer based on the graph from Step 3.
5. Use the combined class data to estimate the probability that a player of this skill level has a streak of at least 5 somewhere in a sequence of 50 free throw attempts.
6. Using the multiplication rule for independent events, you can calculate the probability that a player of this skill level is successful on the *next* 5 free throw attempts:

$$P(SSSSS) = \left(\frac{1}{2}\right)\left(\frac{1}{2}\right)\left(\frac{1}{2}\right)\left(\frac{1}{2}\right)\left(\frac{1}{2}\right) = \left(\frac{1}{2}\right)^5 = 0.031$$

   which is relatively small. At first this might seem inconsistent with your answer in Step 5, but the estimated probability from Step 5 and the calculated probability of 0.031 are really considering different situations. Explain why it is plausible that both probabilities could be correct.
7. Do you think that the assumption that the outcomes of successive free throws are independent is reasonable? Explain. (This is a hotly debated topic among both sports fans and statisticians!)

Copyright 2025 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s). Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

Summary Key Concepts and Formulas

Term or Formula	Comment
Chance experiment	Any experiment for which there is uncertainty concerning the resulting outcome.
Sample space	The collection of all possible outcomes from a chance experiment.
Event	Any collection of possible outcomes from a chance experiment.
Simple event	An event that consists of a single outcome.
Events 1. ***not A***, A^C 2. ***A or B***, $A \cup B$ 3. ***A and B***, $A \cap B$	 1. The event consisting of all outcomes not in *A*. 2. The event consisting of all outcomes in at least one of the two events. 3. The event consisting of outcomes common to both events.
Mutually exclusive events	Events that have no outcomes in common.
Fundamental properties of probability	Fundamental properties of probability 1. The probability of any event must be a number between 0 and 1. 2. If *S* is the sample space for a chance experiment, $P(S) = 1$ 3. If *E* and *F* are mutually exclusive events, $P(E \cup F) = P(E) + P(F)$ 4. $P(E) + P(E^C) = 1$
$P(E) = \dfrac{\text{number of outcomes in } E}{N}$	$P(E)$ when the outcomes are *equally likely* and where *N* is the number of outcomes in the sample space.
$P(E \cup F) = P(E) + P(F)$ $P(E_1 \cup \cdots \cup E_k) = P(E_1) + \cdots + P(E_k)$	Addition rules when events are *mutually exclusive*.
$P(E \mid F) = \dfrac{P(E \cap F)}{P(F)}$	The conditional probability of the event *E* given that the event *F* has occurred.
Independence of events *E* and *F* $P(E \mid F) = P(E)$	Events *E* and *F* are independent if the probability that *E* has occurred given *F* is the same as the probability that *E* will occur with no knowledge of *F*.
$P(E \cap F) = P(E)P(F)$ $P(E_1 \cap \cdots \cap E_k) = P(E_1)P(E_2)\cdots P(E_k)$	Multiplication rules for *independent* events.
$P(E \cup F) = P(E) + P(F) - P(E \cap F)$	The general addition rule for two events.
$P(E \cap F) = P(E \mid F)P(F)$	The general multiplication rule for two events.
$P(E) = P(E \mid B_1)P(B_1) + P(E \mid B_2)P(B_2) + \cdots + P(E \mid B_k)P(B_k)$	The law of total probability, where $B_1, B_2, \ldots, B_k$ are mutually exclusive events with $P(B_1) + P(B_2) + \cdots + P(B_k) = 1$
$P(B_i \mid E) = \dfrac{P(E \mid B_i)P(B_i)}{P(E \mid B_1)P(B_1) + P(E \mid B_2)P(B_2) + \cdots + P(E \mid B_k)P(B_k)}$	Bayes' rule, where $B_1, B_2, \ldots, B_k$ are mutually exclusive events with $P(B_1) + P(B_2) + \cdots + P(B_k) = 1$

Chapter Review Exercises 6.100–6.120

6.100 False positive results are not uncommon with mammograms, a test used to screen for breast cancer. For a patient who has a positive mammogram, the probability that they actually have breast cancer is less than 0.05 if they are under 40 years old, and ranges from 0.05 to 0.109 if they are over 40 years old **("Breast Cancer Screenings: Does the Evidence Support the Recommendations?," Significance [August 2016]: 24–37).** If a patient with a positive mammogram is selected at random, are the two events

B = event that selected patient has breast cancer and

A = event that selected patient is over 40 years old

independent events? Justify your answer using the given information.

6.101 A company uses three different assembly lines—A_1, A_2, and A_3—to manufacture a particular component. Of those manufactured by A_1, 5% need rework to remedy a defect, whereas 8% of A_2's components and 10% of A_3's components need rework. Suppose that 50% of all components are produced by A_1, 30% are produced by A_2 and 20% are produced by A_3.

Copyright 2025 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s). Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

a. Construct a tree diagram with first-generation branches corresponding to the three lines. Leading from each branch, draw one branch for rework (R) and another for no rework (N). Then enter appropriate probabilities on the branches.

b. What is the probability that a randomly selected component came from A_1 and needed rework?

c. What is the probability that a randomly selected component needed rework?

6.102 Consider the following information about passengers on a cruise ship on vacation: 40% check work e-mail, 30% use a cell phone to stay connected to work, 25% bring a laptop with them on vacation, 23% both check work e-mail and use a cell phone to stay connected, and 51% neither check work e-mail nor use a cell phone to stay connected nor bring a laptop. In addition, 88% of those who bring a laptop also check work e-mail and 70% of those who use a cell phone to stay connected also bring a laptop. With

E = event that a traveler on vacation checks work e-mail
C = event that a traveler on vacation uses a cell phone to stay connected
L = event that a traveler on vacation brought a laptop

use the given information to determine the following probabilities. A Venn diagram may help.

a. $P(E)$

b. $P(C)$

c. $P(L)$

d. $P(E \text{ and } C)$

6.103 Use the information given in the previous exercise to determine the following probabilities.

a. $P(E \text{ and } C \text{ or } L)$

b. $P(E|L)$

c. $P(L|C)$

6.104 Use the information given in exercise 6.102 to determine the following probabilities.

a. $P(E \text{ and } C \text{ and } L)$

b. $P(E \text{ and } L)$

c. $P(C \text{ and } L)$

d. $P(C|E \text{ and } L)$

6.105 The article **"Obesity, Smoking Damage U.S. Economy,"** which appeared in the ***Gallup Online Business Journal*** **(gallup.com, September 7, 2016)**, reported that based on a large representative sample of adult Americans, 52.7% claimed that they exercised at least 30 minutes on 3 or more days per week during 2015. It also reported that the percentage for millennials (people age 19–35) was 57.1%, and for those over 35 it was 51.1%. If an adult American were to be selected at random, are the events *selected adult exercises at least 30 minutes 3 times per week* and *selected adult is a millennial* independent or dependent events? Justify your answer using the given information.

6.106 The following table summarizing data on smoking status and age group is consistent with summary quantities obtained in a Gallup Poll published in the online article **"In U.S., Young Adults' Cigarette Use is Down Sharply" (gallup.com, December 10, 2015)**.

	Smoking Status	
Age Group	**Smoker**	**Nonsmoker**
18 to 29	174	618
30 to 49	333	1115
50 to 64	384	1445
65 and older	211	1707

Assume that it is reasonable to consider these data as representative of the American adult population. Consider the chance experiment or randomly selecting an adult American.

a. What is the probability that the selected adult is a smoker?

b. What is the probability that the selected adult is under 50 years of age?

c. What is the probability that the selected adult is a smoker who is 65 or older?

d. What is the probability that the selected adult is a smoker or is age 65 or older?

6.107 A study of the impact of seeking a second opinion about a medical condition is described in the paper **"Evaluation of Outcomes from a National Patient-Initiated Second-Opinion Program" (*The American Journal of Medicine* [2015], 1138e25–1138e33)**. Based on a review of 6791 patient-initiated second opinions, the paper states the following: "Second opinions often resulted in changes in diagnosis (14.8%), treatment (37.4%), or changes in both (10.6%)."

Consider the following two events:

D = event that second opinion results in a change in diagnosis
T = event that second opinion results in a change in treatment

a. What are the values of $P(D)$, $P(T)$, and $P(D \cap T)$?

Copyright 2025 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s). Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

b. What is the probability that a second opinion results in neither a change in diagnosis nor a change in treatment?

c. What is the probability that a second opinion results in a change in diagnosis or a change in treatment?

6.108 A company sends 40% of its overnight mail parcels by means of express mail service A_1. Of these parcels, 2% arrive after the guaranteed delivery time. What is the probability that a randomly selected overnight parcel was shipped by mail service A_1 and was late?

6.109 Return to the context of the previous exercise and suppose that 50% of the overnight parcels are sent by means of express mail service A_2 and the remaining 10% are sent by means of A_3. Of those sent by means of A_2, only 1% arrived late, whereas 5% of the parcels handled by A_3 arrived late.

a. What is the probability that a randomly selected parcel arrived late? (Hint: A tree diagram should help.)

b. Suppose that a randomly selected overnight parcel arrived late. What is the probability that the parcel was shipped using mail service A_1? That is, what is the probability of A_1 given L, denoted $P(A_1|L)$?

c. What is $P(A_2|L)$?

d. What is $P(A_3|L)$?

6.110 Two individuals, A and B, are finalists for a chess championship. They will play a sequence of games, each of which can result in a win for A, a win for B, or a draw. Suppose that the outcomes of successive games are independent, with $P(\text{A wins game}) = 0.3$, $P(\text{B wins game}) = 0.2$, and $P(\text{draw}) = 0.5$. Each time a player wins a game, they earn 1 point and their opponent earns no points. The first player to win 5 points wins the championship.

a. What is the probability that A wins the championship in just five games?

b. What is the probability that it takes just five games to obtain a champion?

c. If a draw earns a half-point for each player, describe how you would perform a simulation to estimate $P(\text{A wins the championship})$. Assume that the championship will end in a draw if both players obtain 5 points at the same time.

6.111 A single-elimination tournament with four players is to be held. In Game 1, the players seeded (rated) first and fourth play. In Game 2, the players seeded second and third play. In Game 3, the winners of Games 1 and 2 play, with the winner of Game 3 declared the tournament winner. Suppose that the following probabilities are given:

$P(\text{seed 1 defeats seed 4}) = 0.8$
$P(\text{seed 1 defeats seed 2}) = 0.6$
$P(\text{seed 1 defeats seed 3}) = 0.7$
$P(\text{seed 2 defeats seed 3}) = 0.6$
$P(\text{seed 2 defeats seed 4}) = 0.7$
$P(\text{seed 3 defeats seed 4}) = 0.6$

a. Describe how you would use random digits to simulate Game 1 of this tournament.

b. Describe how you would use random digits to simulate Game 2 of this tournament.

c. How would you use random digits to simulate Game 3 in the tournament? (This will depend on the outcomes of Games 1 and 2.)

d. Simulate one complete tournament, giving an explanation for each step in the process.

6.112 Refer to the simulation from the previous exercise.

a. Simulate 10 tournaments, and use the resulting information to estimate the probability that the first seed wins the tournament.

b. Ask four classmates for their simulation results. Along with your own results, this should give you information for 50 simulated tournaments. Use this information to estimate the probability that the first seed wins the tournament.

c. Why do the estimated probabilities from Parts (a) and (b) differ? Which do you think is a better estimate of the true probability? Explain.

6.113 In a school machine shop, 60% of all machine breakdowns occur on lathes and 15% occur on drill presses. Let E denote the event that the next machine breakdown is on a lathe, and let F denote the event that a drill press is the next machine to break down. With $P(E) = 0.60$ and $P(F) = 0.15$, calculate:

a. $P(E^C)$

b. $P(E \cup F)$

c. $P(E^C \cap F^C)$

6.114 There are five faculty members in a certain academic department. These individuals have 3, 6, 7, 10, and 14 years of teaching experience. Two of these individuals are randomly selected to serve on a personnel review committee. What is the probability that the chosen representatives have a total of at least 15 years of teaching experience? (Hint: Consider all possible committees.)

Copyright 2025 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s). Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

6.115 The general addition rule for three events states that

$$\begin{aligned} P(A \text{ or } B \text{ or } C) &= P(A) + P(B) + P(C) \\ &\quad - P(A \text{ and } B) - P(A \text{ and } C) \\ &\quad - P(B \text{ and } C) + P(A \text{ and } B \text{ and } C) \end{aligned}$$

A new magazine publishes columns entitled "Art" (A), "Books" (B), and "Cinema" (C). Suppose that

14% of all subscribers read A
23% read B
37% read C
8% read A and B
9% read A and C
13% read B and C
5% read all three columns

What is the probability that a randomly selected subscriber reads at least one of these three columns?

6.116 A theater complex is currently showing four R-rated movies, three PG-13 movies, two PG movies, and one G movie. The following table gives the number of people purchasing tickets for each movie during a week-long period:

Theater	Rating	Number of Viewers
1	R	600
2	PG-13	420
3	PG-13	323
4	R	196
5	G	254
6	PG	179
7	PG-13	114
8	R	205
9	R	139
10	PG	87

Suppose that one of these people is randomly selected.

a. What is the probability that the selected individual saw a PG movie?

b. What is the probability that the selected individual saw a PG or a PG-13 movie?

c. What is the probability that the selected individual did not see an R movie?

6.117 Refer to Exercise 6.116, but now suppose that two viewers are randomly selected (without replacement). Let R_1 and R_2 denote the events that the first and second individuals, respectively, watched an R-rated movie. Are R_1 and R_2 independent events? Explain. From a practical point of view, can these events be regarded as independent? Explain.

6.118 Suppose that a box contains 25 light bulbs, of which 20 are good and the other 5 are defective. Consider randomly selecting three bulbs without replacement. Let E denote the event that the first bulb selected is good, F be the event that the second bulb is good, and G represent the event that the third bulb selected is good.

a. What is $P(E)$?

b. What is $P(F|E)$?

c. What is $P(G|E \cap F)$?

d. What is the probability that all three selected bulbs are good?

6.119 Return to Exercise 6.118, and suppose that 4 bulbs are randomly selected from the 25.

a. What is the probability that all 4 are good?

b. What is the probability that at least 1 selected bulb is bad?

6.120 A transmitter is sending a message using a binary code (a sequence of 0's and 1's). Each transmitted bit (0 or 1) must pass through three relays to reach the receiver. At each relay, the probability is 0.20 that the bit sent on is different from the bit received (a reversal). Assume that the relays operate independently of one another:

transmitter $\rightarrow$ relay 1 $\rightarrow$ relay 2 $\rightarrow$ relay 3 $\rightarrow$ receiver

a. If a 1 is sent from the transmitter, what is the probability that a 1 is sent on by all three relays?

b. If a 1 is sent from the transmitter, what is the probability that a 1 is received by the receiver? (Hint: The eight experimental outcomes can be displayed on a tree diagram with three generations of branches, one generation for each relay.)

Chapter 7

Random Variables and Probability Distributions

Erika Cristina Manno/Shutterstock.com

This chapter is the first of two chapters that together link the basic ideas of probability introduced in Chapter 6 with the methods of statistical inference. Chapter 6 used probability to describe the long-run relative frequency of occurrence of various types of outcomes. In this chapter probability models that can be used to describe the distribution of values of a variable are introduced.

In Chapter 8, you will see how these same probability models can be used to describe the behavior of sample statistics. You will also see that this is what allows you to draw conclusions based on sample data.

In a chance experiment, you often focus on some numerical aspect of the outcome. For example, an environmental scientist who obtains an air sample from a specified location might be interested in the ozone concentration (ozone is a major constituent of air pollution). A quality control inspector who must decide whether to accept a large shipment of plastic cell phone cases may base the decision on the number of defective cases in a group of 20 cases randomly selected from the shipment.

Before selecting an air sample, the value of the ozone concentration is uncertain. Similarly, the number of defective cases among the 20 selected for inspection might be any whole number between 0 and 20. Because the value of a variable quantity such as ozone concentration or number of defective cases is subject to uncertainty, such variables are called *random variables*.

This chapter begins by distinguishing between discrete and continuous numerical random variables. You will see how the behavior of both discrete and continuous numerical random variables can be described by a probability distribution. This distribution can then be used to make probability statements about values of the random variable. Three commonly encountered probability distributions (the binomial, geometric, and normal distributions) are also introduced.

Learning Objectives

After completing this chapter, you will be able to:

- Explain how a probability distribution describes the long-run behavior of a random variable.
- Distinguish between discrete and continuous random variables.

Copyright 2025 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s). Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

- Construct the probability distribution of a discrete random variable.
- Calculate and interpret the mean and standard deviation of a discrete random variable.
- Distinguish between binominal and geometric random variables.
- Calculate and interpret binomial probabilities.
- Interpret areas under the density curve for a continuous random variable as probabilities.
- Calculate probabilities involving continuous random variables whose density curves have a simple form.
- Interpret areas under a normal curve as probabilities.
- Identify extreme values for a random variable that has a normal distribution.
- Construct and interpret a normal probability plot.

Section 7.1 Random Variables

In many chance experiments, interest may focus on one or more variable quantities. For example, consider a management consultant who is studying the operation of a grocery store. A chance experiment might involve randomly selecting a customer leaving the store. One numerical variable that might be of interest is the number of items purchased by the customer. This variable can be represented using a letter, such as x. Possible values of this variable are 0, 1, 2, 3, and so on. The possible values of x are isolated points on the number line. Until a customer is selected and the number of items counted, there is uncertainty about what value of x will be observed.

Another variable of interest might be the time y (in minutes) spent in a checkout line. One possible value of y is 3.0 minutes and another is 4.0 minutes, but *any* other number between 3.0 and 4.0 is also a possibility. The possible y values form an entire interval (a continuum) on the number line.

Definitions

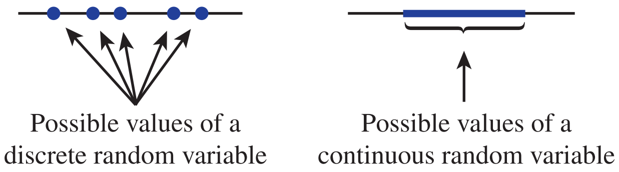
Random variable: A numerical variable whose value depends on the outcome of a chance experiment. A random variable associates a numerical value with each outcome of a chance experiment.

Discrete random variable: A random variable is discrete if its set of possible values is a collection of isolated points along the number line.

Continuous random variable: A random variable is continuous if its set of possible values includes an entire interval on the number line.

In this text, lowercase letters, such as x and y, are used to represent random variables.* Figure 7.1 shows a set of possible values for each type of random variable. In practice, a discrete random variable is usually the result of counting (for example, the number of items purchased, the number of gas pumps in use, or the number of broken eggs in a carton). A continuous random variable is one whose value is typically obtained by measurement (for example, temperature in a freezer, weight of a pineapple, or amount of time spent in a store).

Figure 7.1
Two different types of random variables.


Possible values of a discrete random variable
Possible values of a continuous random variable

---

*In some books, uppercase letters are used to name random variables, with lowercase letters representing a particular value that the variable might assume. We have chosen to use a simpler and less formal notation.

Because there is a limit to the accuracy of any measuring instrument, such as a watch or a scale, it may seem that any variable should be regarded as discrete. For example, when weight is measured to the nearest pound, the observed values appear to be isolated points along the number line, such as 2 pounds or 3 pounds. But this is just a function of the accuracy with which weight is recorded and not because a weight between 2 and 3 pounds is impossible. In this case, the variable is continuous.

Example 7.1 Book Sales

Consider an experiment in which the type of book, print (P) or digital (D), chosen by each of three successive customers making a purchase of a single book from on an online bookstore is noted. Define a random variable x by

x = number of customers purchasing a digital book

The outcome in which the first and third customers purchase a digital book and the second customer purchases a print book can be abbreviated DPD. The associated x value is 2, because two of the three customers selected a digital book. Similarly, the x value for the outcome DDD (all three purchase a digital book) is 3.

The eight possible outcomes and the corresponding values of x are displayed in the following table:

Outcome	PPP	DPP	PDP	PPD	DDP	DPD	PDD	DDD
x value	0	1	1	1	2	2	2	3

There are only four possible x values—0, 1, 2, and 3—and these are isolated points on the number line. This means that x is a discrete random variable.

In some situations, the random variable of interest is discrete, but the number of possible values is not finite. This is illustrated in Example 7.2.

Example 7.2 This Could Be a Long Game . . .

Two friends agree to play a game that consists of a sequence of trials. The game continues until one player wins two trials in a row. One random variable of interest might be

x = number of trials required to complete the game

Let A denote a win for Player 1 and B denote a win for Player 2. The simplest possible experimental outcomes are AA (the case in which Player 1 wins the first two trials and the game ends) and BB (the case in which Player 2 wins the first two trials). With either of these two outcomes, $x = 2$. There are also two outcomes for which $x = 3$: ABB and BAA. Some other possible outcomes and associated x values are shown in the following table.

Outcomes	x value
AA, BB	2
BAA, ABB	3
ABAA, BABB	4
ABABB, BABAA	5
⋮	⋮
ABABABABAA, BABABABABB	10

Notice that any positive integer that is 2 or greater is a possible value. Because the values 2, 3, 4, … are isolated points on the number line, x is a discrete random variable even though there is no upper limit for the number of possible values.

Copyright 2025 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s). Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

Example 7.3 Stress

In an engineering stress test, pressure is applied to a thin 1-foot-long bar until the bar snaps. The precise location where the bar will snap is uncertain. Let x be the distance from the left end of the bar to the break. Then $x = 0.25$ is one possibility, $x = 0.9$ is another, and any other number between 0 and 1 is also a possible value of x. (Figure 7.2 shows the case of the outcome $x = 0.6$.) The set of possible values is an entire interval on the number line, so x is a continuous random variable.

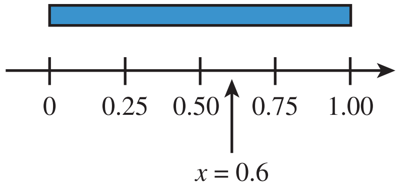


Figure 7.2
The bar for Example 7.3 and the outcome $x = 0.6$.

Even though in practice you may only be able to measure the distance to the nearest tenth of an inch or hundredth of an inch, the *actual* distance could be any number between 0 and 1. So, even though the recorded values might be rounded because of the accuracy of the measuring instrument, the variable is still continuous.

In data analysis, random variables often arise in the context of summarizing sample data when a sample is selected from some population. This is illustrated in Example 7.4.

Example 7.4 College Plans

Suppose that a high school counselor plans to select a random sample of 50 seniors and to ask each student in the sample whether they plan to attend college after graduation. The process of sampling is a chance experiment. The sample space for this experiment consists of all the different possible random samples of size 50 that might result (there are a very large number of these), and for simple random sampling, each of these outcomes is equally likely.

Suppose that the counselor is interested in the number of students in the sample who plan to attend college. The letter x can be used to represent this number:

x = number of students in the sample who plan to attend college

Then x is a random variable, because it associates a numerical value with each of the possible outcomes (random samples). Possible values of x are 0, 1, 2, . . . , 50, and x is a discrete random variable.

Exercises 7.1–7.7

7.1 State whether each of the following random variables is discrete or continuous:
- **a.** The number of defective tires on a car
- **b.** The body temperature of a hospital patient
- **c.** The number of pages in a book
- **d.** The number of draws (with replacement) from a deck of playing cards until a heart is selected
- **e.** The lifetime of a lightbulb

7.2 Classify each of the following random variables as either discrete or continuous:
- **a.** The fuel efficiency (mpg) of an automobile
- **b.** The amount of rainfall at a particular location during the next year
- **c.** The distance that a person throws a baseball
- **d.** The number of questions asked during a 1-hour lecture
- **e.** The weight of a 2-month-old golden retriever puppy
- **f.** The amount of water used by a household during a given month
- **g.** The number of traffic citations issued by the highway patrol in a particular county on a given day

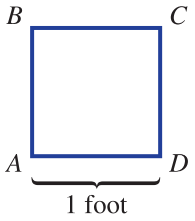
Copyright 2025 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s). Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

7.3 Starting at a particular time, each car entering an intersection is observed to see whether it turns left (L) or right (R) or goes straight ahead (S). The experiment terminates as soon as a car is observed to go straight. Suppose that the random variable y denotes the number of cars observed.

a. What are possible y values?

b. List five different outcomes and their associated y values. (Hint: See Example 7.2.)

7.4 A point is randomly selected from the interior of the square pictured.



Let x denote the distance from the lower left-hand corner A of the square to the selected point.

a. What are possible values of x?

b. Is x a discrete or a continuous variable? (Hint: See Example 7.3.)

7.5 A point is randomly selected on the surface of a lake that has a maximum depth of 100 feet. Let y be the depth of the lake at the randomly chosen point.

a. What are possible values of y?

b. Is y discrete or continuous?

7.6 A person stands at the corner marked A of the square pictured in Exercise 7.4 and tosses a coin. If it lands heads up, the person moves one corner clockwise, to B. If the coin lands tails up, the person moves one corner counterclockwise, to D. This process is then repeated until the person arrives back at A. Let y denote the number of coin tosses.

a. What are possible values of y?

b. Is y discrete or continuous?

7.7 A box contains four slips of paper marked 1, 2, 3, and 4. Two slips are selected without replacement. List the possible values for each of the following random variables:

a. x = sum of the two numbers

b. y = difference between the first and second numbers

c. z = number of slips selected that show an even number

d. w = number of slips selected that show a 4

Section 7.2 Probability Distributions for Discrete Random Variables

The probability distribution for a random variable describes the long-run behavior of the variable. For example, suppose that the county Department of Animal Regulation is interested in studying the variable

x = number of licensed dogs or cats in a household

County regulations prohibit more than five dogs or cats per household. If you consider the chance experiment of randomly selecting a household in this county, then x is a discrete random variable because it associates a numerical value (0, 1, 2, 3, 4, or 5) with each of the possible outcomes (households) in the sample space.

Although you know what the possible values for x are, it would also be useful to know how this variable behaves in repeated observation. What would be the most common value? What proportion of the time would $x = 5$ be observed? $x = 3$? A probability distribution provides this type of information about the long-run behavior of a random variable.

The **probability distribution of a discrete random variable *x*** gives the probability associated with each possible x value. Each probability is the long-run relative frequency of occurrence of the corresponding x value when a chance experiment is performed a very large number of times.

Common ways to define a probability distribution for a discrete random variable are a table, a probability histogram, or a formula.

Notation: If one possible value of x is 2, $p(2)$ is often used place of $P(x = 2)$. Similarly, $p(5)$ denotes the probability that $x = 5$, and so on.

Copyright 2025 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s). Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

Example 7.5 Electric and Hybrid Cars

Understand the context ➤

The ***Washington Post*** **(August 6, 2021)** reported that about 7% of new car purchases were electric or hybrid cars rather than gas-powered cars. Suppose that each of four randomly selected customers purchasing a new car at a large dealership chooses either an electric or hybrid car (E) or select a gas-powered car (G). Assume that these customers make their choices independently of one another and that 7% of all customers select an electric or hybrid car. This implies that for any particular one of the four customers, $P(E) = 0.07$ and $P(G) = 0.93$. One possible outcome is $EGGE$, where the first and fourth customers select electric or hybrid models and the other two choose gas-powered models. Because the customers make their choices independently, the multiplication rule for independent events implies that

$$\begin{aligned} P(EGGE) &= P(\text{1st chooses } E \textit{ and } \text{2nd chooses } G \textit{ and } \text{3rd chooses } G \textit{ and } \text{4th chooses } E) \\ &= P(E)P(G)P(G)P(E) \\ &= (0.07)(0.93)(0.93)(0.07) \\ &= 0.00424 \end{aligned}$$

Similarly,

$$\begin{aligned} P(EGEG) &= P(E)P(G)P(E)P(G) \\ &= (0.07)(0.93)(0.07)(0.93) \\ &= 0.00424 \text{ (which is equal to } P(EGGE)) \end{aligned}$$

and

$$P(GGGE) = (0.93)(0.93)(0.93)(0.07) = 0.05630$$

The number among the four customers who purchase electric or hybrid models is a random variable. Using x to denote this random variable,

x = number of electric or hybrid cars purchased by a group of four customers

Table 7.1 displays the 16 possible outcomes that might be observed, the probability of each outcome, and the value of the random variable x that is associated with each outcome.

Table 7.1 Outcomes and Probabilities for Example 7.5

Outcome	Probability	x Value	Outcome	Probability	x Value
$GGGG$	0.74805	0	$GEEG$	0.00424	2
$EGGG$	0.05630	1	$GEGE$	0.00424	2
$GEGG$	0.05630	1	$GGEE$	0.00424	2
$GGEG$	0.05630	1	$GEEE$	0.00032	3
$GGGE$	0.05630	1	$EGEE$	0.00032	3
$EEGG$	0.00424	2	$EEGE$	0.00032	3
$EGEG$	0.00424	2	$EEEG$	0.00032	3
$EGGE$	0.00424	2	$EEEE$	0.00002	4

Do the work ➤

The probability distribution of x is easily obtained from this information. Consider the smallest possible x value, 0. The only outcome for which $x = 0$ is $GGGG$, so

$$p(0) = P(x = 0) = P(GGGG) = 0.74805$$

There are four different outcomes for which $x = 1$, so $p(1)$ results from summing the four corresponding probabilities:

$$\begin{aligned} p(1) = P(x = 1) &= P(EGGG \textit{ or } GEGG \textit{ or } GGEG \textit{ or } GGGE) \\ &= P(EGGG) + P(GEGG) + P(GGEG) + P(GGGE) \\ &= 0.05630 + 0.05630 + 0.05630 + 0.05630 \\ &= 4(0.05630) \\ &= 0.22520 \end{aligned}$$

Copyright 2025 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s). Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

Similarly,

$$p(2) = P(EEGG) + \cdots + P(GGEE) = 6(0.00424) = 0.02544$$
$$p(3) = 4(0.00032) = 0.00128$$
$$p(4) = 0.00002$$

The probability distribution of x is summarized in the following table:

x Value	0	1	2	3	4
$p(x)$	0.74805	0.22520	0.02544	0.00128	0.00002

Interpret the results ➤

To interpret $p(1) = 0.22520$, think of performing the chance experiment repeatedly, each time with a new group of four customers. In the long run, 22.52% of these groups of four will have exactly one customer who purchases an electric or hybrid car.

The probability distribution can be used to determine probabilities of various events involving x. For example, the probability that at least two of the four customers choose electric or hybrid models is

$$\begin{aligned} P(x \geq 2) &= P(x = 2 \text{ } or \text{ } x = 3 \text{ } or \text{ } x = 4) \\ &= p(2) + p(3) + p(4) \\ &= 0.02674 \end{aligned}$$

This means that in the long run, a group of four people purchasing cars will include at least two who select electric or hybrid models about 2.6% of the time.

A probability distribution table for a discrete variable shows the possible x values and also $p(x)$ for each possible x value. Because $p(x)$ is a probability, it must be a number between 0 and 1, and because the probability distribution includes all possible x values, the sum of all the $p(x)$ values must equal 1. These properties of discrete probability distributions are summarized in the following box.

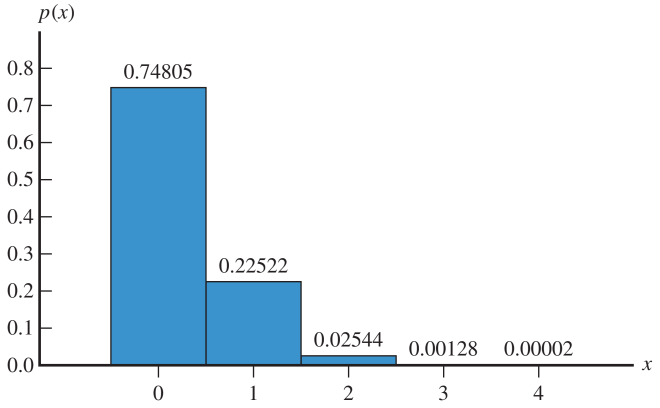
Properties of Discrete Probability Distributions

Properties of Discrete Probability Distributions

1. For every possible x value, $0 \leq p(x) \leq 1$.
2. $\sum_{\text{all } x \text{ values}} p(x) = 1$

A graphical representation of a discrete probability distribution is called a *probability histogram*. The probability histogram has a rectangle centered above each possible value of x, and the area of each rectangle is proportional to the probability of the corresponding value. Figure 7.3 displays the probability histogram for the probability distribution of Example 7.5.

Figure 7.3
Probability histogram for the distribution of Example 7.5.



Copyright 2025 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s). Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

In Example 7.5, the probability distribution was constructed by starting with a chance experiment and using probability rules. When it is not possible to construct a probability distribution in this way, an investigator can sometimes propose an approximate probability distribution consistent with observed values of the random variable and prior knowledge. Probability distributions based on observed values must still be consistent with rules of probability:

1. $p(x) \geq 0$ for every x value.
2. $\sum_{\text{all } x \text{ values}} p(x) = 1$

This is illustrated in Examples 7.6 and 7.7.

Example 7.6 Paint Flaws

Understand the context ➤

In automobile manufacturing, one of the last steps in the process of assembling a new car is painting. Some minor blemishes in the paint surface are considered acceptable, but if there are too many, it becomes noticeable to a potential customer and the car must be repainted. Cars coming off the assembly line are carefully inspected and the number of minor blemishes in the paint surface is determined. Consider the random variable x to be the number of minor blemishes on a randomly selected car from a particular manufacturing plant. A large number of automobiles were evaluated, and a probability distribution consistent with these observations is

x	0	1	2	3	4	5	6	7	8	9	10
$p(x)$	0.041	0.010	0.209	0.223	0.178	0.114	0.061	0.028	0.011	0.004	0.001

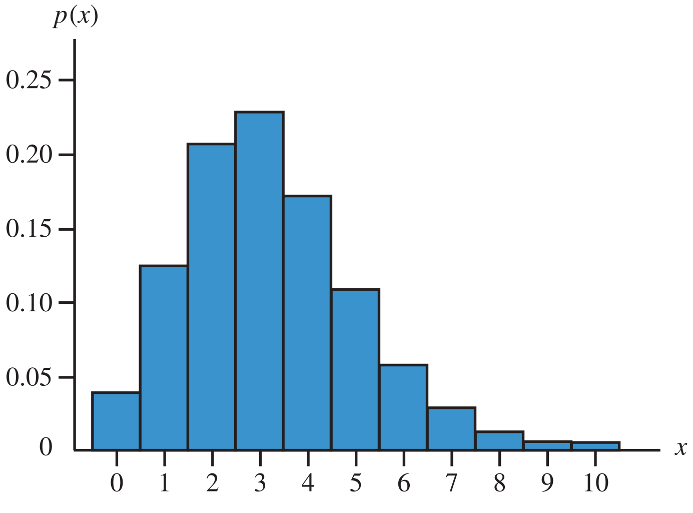
Interpret the results ➤

The corresponding probability histogram appears in Figure 7.4. The probabilities in this distribution reflect the car manufacturer's experience. For example, $p(3) = 0.223$ indicates that 22.3% of new automobiles have 3 minor paint blemishes. The probability that the number of minor paint blemishes is between 2 and 5 inclusive is

$$P(2 \leq x \leq 5) = p(2) + p(3) + p(4) + p(5) = 0.724$$

If car after car of this type were examined, in the long run, 72.4% would have 2, 3, 4, or 5 minor paint blemishes.

Figure 7.4
Probability histogram for the distribution of the number of minor paint blemishes on a randomly selected car.



Example 7.7 Shuffling Music

Understand the context ➤

Many people use apps like Spotify and Pandora to listen to music on a smartphone or a computer. Most of these apps have a shuffle feature that takes a group of songs and plays them in a random order. Some users of these apps have questioned the "randomness" of the shuffle, saying that they often hear several songs from the same artist played near each other. A blog post from Spotify titled **"How to Shuffle Songs"** says

Copyright 2025 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s). Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

> At Spotify we take user feedback seriously. We noticed some users complaining about our shuffling algorithm playing a few songs from the same artist right after each other. The users were asking “Why isn’t your shuffling algorithm random?” We responded “Hey! Our shuffling is random!”

Are several songs from the same artist playing one after another consistent with a “random” shuffle? To investigate the claim that a shuffle that often plays songs from the same artist back to back is not consistent with random shuffling, you can start by creating a playlist. For example, consider a playlist that was formed by adding the following to the list:

Artist A: an album consisting of 10 songs
Artist B: an album consisting of 9 songs
Artist C: an album consisting of 6 songs
Artist D: 4 songs
Artist E: 5 songs
Artists F, G, H, I, J, and K: 1 song each

This playlist consists of 40 songs from 11 different artists.

What Do You Think?

For a playlist like this one (40 songs from 11 different artists), do you think it would be surprising to hear two songs by the same artist in a row in a random shuffle? Would you be surprised to hear four or more songs in a row by the same artist?

Do the work ➤

To investigate, you could carry out a simulation by creating a random shuffle of these 40 songs and looking to see if there are songs from the same artist played back to back in the shuffle. A statistics software package or app could be used create a random shuffle of the 40 songs. For example, a data set with 10 A’s, 9 B’s, 6 C’s, 4 D’s, 5 E’s, and one each of F, G, H, I, J, and K was created in Minitab and the random shuffle shown here was created:

Song Number	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
Artist	E	D	I	C	J	E	**A**	**A**	**B**	**B**	A	E	A	E	A	D	C	**B**	**B**	**B**
Song Number	21	22	23	24	25	26	27	28	29	30	31	32	33	34	35	36	37	38	39	40
Artist	A	**C**	**C**	E	**B**	**B**	**B**	A	C	D	F	G	**A**	**A**	K	B	D	C	A	H

Cases where songs by the same artist appear together **have been highlighted**. Notice that in this random shuffle, two songs by the same artist were played back to back four times and three songs by the same artist were played in a row two times. Even though this was a random shuffle, there were quite a few times where songs from the same artist were played consecutively. For this random shuffle, the length of the longest run of songs by the same artist is 3.

Is this typical, or was this just an unlucky shuffle? Could you see even longer runs in random shuffles? Repeating the process of creating a shuffle that you know is random many times and looking at the length of the longest run by the same artist for each of the shuffles would enable you to estimate the probabilities needed for the distribution of

x = length of the longest run of songs by the same artist

for a playlist like the one created for this example. The smallest possible value for x is 1, which would correspond to a shuffle where there is never a situation where the same artist is played twice in a row. The largest possible value for x is 10, which is the largest number of songs by any one artist and corresponds to a shuffle where all 10 songs by artist A were played in a row.

Minitab was used to create 100 random shuffles and these shuffles were then used to calculate the approximate probabilities in the accompanying table.

x	1	2	3	4	5	6	7	8	9	10
$p(x)$	0	0.50	0.39	0.06	0.04	0	0	0.01	0	0

Copyright 2025 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s). Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

The probabilities in the table were just estimated from 100 simulated observations of x, the probabilities that are shown as 0 in the table are not actually equal to 0. They are just small enough that these values were not observed in 100 random shuffles.

Interpret the results ➤ Based on the estimated probability distribution, you can see that it is not unusual to have several songs by the same artist playing in a row for a playlist like the one created in this example. However, many people misinterpreted this as a flaw in the random shuffle. The "How to Shuffle Songs" blog post quoted earlier in this example went on to say

> At first, we didn't understand what the users were trying to tell us by saying that the shuffling is not random, but then we read the comments more carefully and noticed that some people don't want the same artist playing two or three times within a short time period. … so we decided to look into ways of changing our shuffling algorithm so that the users are happier. We learned that they don't like perfect randomness.

Now, many of the music apps don't use a truly random shuffle but the shuffle now behaves more like what users mistakenly think a random shuffle should!

You have seen examples in which the probability distribution of a discrete random variable has been given as a table or as a probability histogram. It is also possible to give a formula that allows calculation of the probability for each possible value of the random variable. Examples of this approach are given in Section 7.5.

Exercises 7.8–7.21

7.8 Define the random variable x to be the number of courses for which a randomly selected student at a certain university is registered. Suppose that the probability distribution of x is given in the following table:

x	1	2	3	4	5	6	7
$p(x)$	0.02	0.03	0.09	0.25	0.40	0.16	0.05

a. What is $P(x = 4)$?
b. What is $P(x \leq 4)$?
c. What is the probability that the selected student is taking at most five courses?
d. What is the probability that the selected student is taking at least five courses?
e. What is the probability that the selected student is taking more than five courses?

7.9 Using the probability distribution given in the previous exercise, calculate $P(3 \leq x \leq 6)$ and $P(3 < x < 6)$. Explain in words why these two probabilities are different.

7.10 Let y denote the number of broken eggs in a randomly selected carton of one dozen eggs. Suppose that the probability distribution of y is as follows:

y	0	1	2	3	4
$p(y)$	0.65	0.20	0.10	0.04	?

a. Only y values of 0, 1, 2, 3, and 4 have positive probabilities. What is $p(4)$? (Hint: Consider the properties of a discrete probability distribution.)
b. How would you interpret $p(1) = 0.20$?
c. Calculate $P(y \leq 2)$, the probability that the carton contains at most two broken eggs, and interpret this probability.
d. Calculate $P(y < 2)$, the probability that the carton contains *fewer than* two broken eggs. Why is this smaller than the probability in Part (c)?

7.11 Use the probability distribution given in the previous exercise to answer the following questions.
a. What is the probability that the carton contains exactly 10 unbroken eggs?
b. What is the probability that at least 10 eggs are unbroken?

7.12 Suppose that fund-raisers at a university call recent graduates to request donations for campus outreach programs. They report the following information for last year's graduates:

Size of donation	\$0	\$10	\$25	\$50
Proportion of calls	0.45	0.30	0.20	0.05

Three attempts were made to contact each graduate. A donation of \$0 was recorded both for those who were contacted but who declined to make a donation and for those who were not reached in three attempts. Consider the variable x = amount of donation for a person selected at random from the population of last year's graduates of this university.
a. Write a few sentences describing what you think you might see if the value of x was observed for each of 1000 graduates.
b. In the long run, what value of x would be observed most often?
c. What is $P(x \geq 25)$?
d. What is $P(x > 0)$?

7.13 Airlines sometimes overbook flights. Suppose that for a plane with 100 seats, an airline takes 110 reservations. Define the variable x as the

Copyright 2025 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s). Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

number of people who actually show up for a sold-out flight. From past experience, the probability distribution of x is given in the following table:

x	95	96	97	98	99	100	101	102
$p(x)$	0.05	0.10	0.12	0.14	0.24	0.17	0.06	0.04

x	103	104	105	106	107	108	109	110
$p(x)$	0.03	0.02	0.01	0.005	0.005	0.005	0.0037	0.0013

a. What is the probability that the airline can accommodate everyone who shows up for the flight?
b. What is the probability that not all passengers can be accommodated?
c. If you are trying to get a seat on such a flight and you are number 1 on the standby list, what is the probability that you will be able to take the flight? What if you are number 3?

7.14 Suppose that a computer manufacturer receives computer boards in lots of five. Two boards are selected from each lot for inspection. You can represent possible outcomes of the selection process by pairs. For example, the pair (1,2) represents the outcome where Boards 1 and 2 are selected for inspection.
a. List the 10 different possible outcomes.
b. Suppose that Boards 1 and 2 are the only defective boards in a lot of five. Two boards are to be chosen at random. Define x to be the number of defective boards observed among those inspected. Find the probability distribution of x.

7.15 Simulate the chance experiment described in the previous exercise using five slips of paper, with two marked *defective* and three marked *not defective*. Place the slips in a box, mix them well, and draw out two. Record the number of defective boards. Replace the slips and repeat until you have 50 observations of the value of x. Construct a relative frequency distribution for the 50 observations, and compare this with the probability distribution obtained in the previous exercise.

7.16 Of all airline flight requests received by a certain discount ticket broker, 70% are for domestic travel (D) and 30% are for international flights (I). Define the random variable x to be the number of requests among the next three that are for domestic flights. Assuming independence of successive requests, determine the probability distribution of x. (Hint: See Example 7.5.)

7.17 Suppose that 20% of all homeowners in an earthquake-prone area of California are insured against earthquake damage. Four homeowners are selected at random. Suppose that x denotes the number among the four who have earthquake insurance.
a. Find the probability distribution of x. (Hint: Let S represent a homeowner who has insurance and F one who does not. Then one possible outcome is SFSS, with probability (0.2)(0.8)(0.2)(0.2) and associated x value of 3. There are 15 other outcomes.)
b. What is the most likely value of x?
c. What is the probability that at least two of the four selected homeowners have earthquake insurance?

7.18 A box contains five slips of paper, marked \$1, \$1, \$1, \$10, and \$25. The winner of a contest selects two slips of paper at random and then gets the larger of the dollar amounts on the two slips. Define a random variable w by w = amount awarded. Determine the probability distribution of w. (Hint: Think of the slips as numbered 1, 2, 3, 4, and 5, so that an outcome of the experiment consists of two of these numbers.)

7.19 Components coming off an assembly line are either free of defects (S, for success) or defective (F, for failure). Suppose that 70% of the components are defect-free. Components are independently selected and tested one by one. Let y denote the number of components that must be tested until a defect-free component is obtained.
a. What is the smallest possible y value, and what outcome corresponds to this y value? What is the second smallest y value, and what is the corresponding outcome?
b. What is the set of all possible y values?
c. Calculate the probability of each of the five smallest y values. (You should see a pattern that leads to a simple formula for $p(y)$, the probability distribution of y.)

7.20 When applying for a building permit, a contractor is required by a county planning department to submit anywhere from one to five forms (depending on the nature of the project). Define the random variable y to be the number of forms required of the next applicant. Suppose that the probability that y forms are required is known to be proportional to y with $p(y) = ky$ for $y = 1, \ldots, 5$.
a. What is the value of k? (Hint: $\Sigma p(y) = 1$.)
b. What is the probability that at most three forms are required?
c. What is the probability that between two and four forms (inclusive) are required?
d. Could $p(y) = y^2/50$ for $y = 1, 2, 3, 4, 5$ be the probability distribution of y? Explain.

7.21 A new battery's voltage may be acceptable (A) or unacceptable (U). A certain flashlight requires two batteries, so batteries will be independently selected and tested until two acceptable ones have been found. Suppose that 80% of all batteries have acceptable voltages, and let y = number of batteries that must be tested.
a. What are the possible values of y?
b. What is $p(2) = P(y = 2)$?
c. What is $p(3)$? (Hint: There are two different outcomes that result in $y = 3$.)

Copyright 2025 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s). Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

Section 7.3 Probability Distributions for Continuous Random Variables

A continuous random variable is one that has possible values that form an entire interval on the number line. One example of a continuous random variable is the weight *x* (in pounds) of a full-term newborn baby. Suppose for the moment that weight is recorded only to the nearest pound. Then a reported weight of 7 pounds would be used for any weight greater than or equal to 6.5 pounds and less than 7.5 pounds.

The probability distribution of *x* can be pictured as a probability histogram with rectangles centered at 4, 5, and so on. The area of each rectangle represents the probability of the corresponding weight value, and the total area of all the rectangles is 1. The probability that a weight (to the nearest pound) is between two values, such as 6 and 8, is the sum of the areas of the corresponding rectangles. Figure 7.5(a) illustrates this.

Now suppose that weight is measured to the nearest tenth of a pound. There are many more possible reported weight values than before, such as 5.0, 5.1, 5.7, 7.3, and 8.9. As shown in Figure 7.5(b), the rectangles in the probability histogram are much narrower, and this histogram has a much smoother appearance.

Now think about what would happen if weight were measured with greater accuracy. The probability histogram will look more like the smooth curve shown in Figure 7.5(c). This curve does not go below the horizontal measurement scale, and the total area under the curve is 1 (because this is true for all probability histograms). The probability that *x* falls in an interval such as $6 \le x \le 8$ is represented by the area under the curve and above that interval.

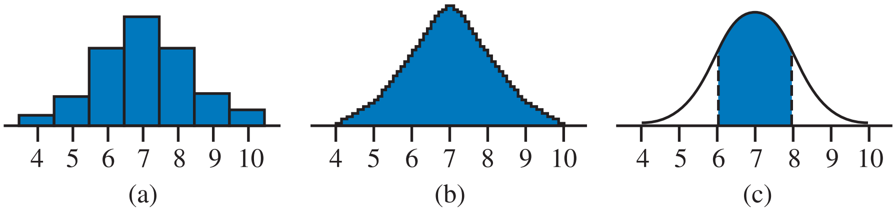


Figure 7.5
Probability distribution for birth weight: (a) weight measured to the nearest pound; (b) weight measured to the nearest tenth of a pound; (c) limiting curve as measurement accuracy increases; shaded area = $P(6 \le \text{weight} \le 8)$.

A **probability distribution for a continuous random variable *x*** is specified by a curve called a **density curve**. The function that defines this curve is denoted by $f(x)$ and it is called the **density function**.

The following are properties of all continuous probability distributions:

1. $f(x) \ge 0$ (the curve cannot dip below the horizontal axis).
2. The total area under the density curve is equal to 1.

The probability that *x* falls in any particular interval is equal to the area under the density curve and above the interval.

Many probability calculations for continuous random variables involve the following three types of events:

1. $a < x < b$, the event that the random variable *x* assumes a value between two given numbers, *a* and *b*
2. $x < a$, the event that the random variable *x* assumes a value less than a given number *a*
3. $x > b$, the event that the random variable *x* assumes a value greater than a given number *b* (this can also be written as $b < x$)

Figure 7.6 illustrates how the probabilities of these events are represented by areas under a density curve.

Copyright 2025 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s). Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

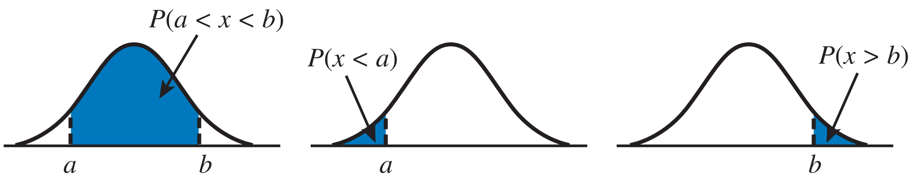


Figure 7.6
Probabilities as areas under a density curve.

Example 7.8 Application Processing Times

Understand the context ➤ Suppose x represents the continuous random variable

x = amount of time (in minutes) it takes to process a credit card application

Suppose that x has a probability distribution with density function

$$f(x) = \begin{cases} 0.5 & 4 < x < 6 \\ 0 & \text{otherwise} \end{cases}$$

The graph of the density curve $f(x)$ is shown in Figure 7.7(a). It is easy to use this density curve to calculate probabilities, because areas under this density curve are areas of rectangles. The area of a rectangle is calculated using the formula

$$\text{Area} = (\text{Base})(\text{Height})$$

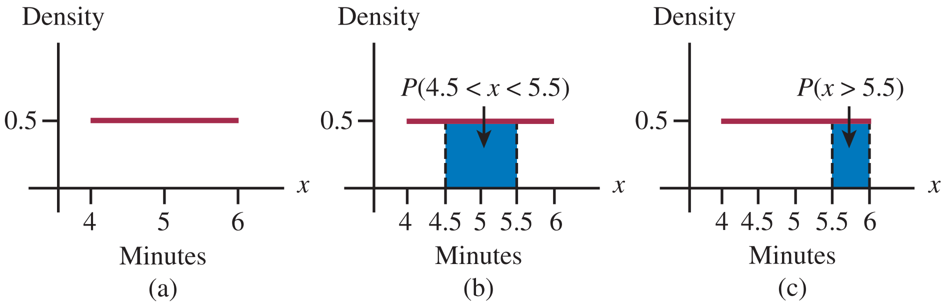


Figure 7.7
The probability distribution for Example 7.8.

Do the work ➤ The curve has positive height, 0.5, only between $x = 4$ and $x = 6$. The total area under the curve is just the area of the rectangle with base extending from 4 to 6 and with height 0.5. You can verify that the total area under the density curve is equal to 1 by calculating

$$\text{Area} = (6 - 4)(0.5) = 1$$

When the density is constant over an interval (resulting in a horizontal density curve), the probability distribution is called a **uniform distribution**.

As illustrated in Figure 7.7(b), the probability that x is between 4.5 and 5.5 is

$$\begin{aligned} P(4.5 < x < 5.5) &= \text{Area of shaded rectangle} \\ &= (\text{Base})(\text{Height}) \\ &= (5.5 - 4.5)(0.5) \\ &= 0.5 \end{aligned}$$

Similarly (see Figure 7.7(c)), because in this context $x > 5.5$ is equivalent to $5.5 < x \leq 6$,

$$P(x > 5.5) = (6 - 5.5)(0.5) = 0.25$$

Inerpret the results ➤ Based on this probability distribution, in the long run, 25% of all applications will have processing times that are greater than 5.5 minutes.

The probability that a *discrete* random variable x takes a value in an interval between two endpoints a and b can depend on whether either endpoint is included in the interval. For example, suppose that x is the number of major defects on a new car. Then

$$P(3 \leq x \leq 7) = p(3) + p(4) + p(5) + p(6) + p(7)$$

Copyright 2025 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s). Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

and

$$P(3 < x < 7) = p(4) + p(5) + p(6)$$

However, if x is a *continuous* random variable, such as task completion time, then

$$P(3 \leq x \leq 7) = P(3 < x < 7)$$

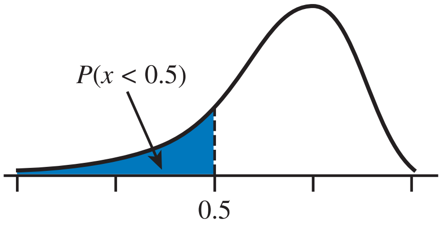
because the area under a density curve and above a single value such as 3 or 7 is 0. Geometrically, you can think of finding the area above a single point as finding the area of a rectangle with width = 0. This means that for a continuous random variable, the area above an interval of values does not depend on whether either endpoint is included.

If x is a continuous random variable, then for any two numbers a and b with $a < b$,

$$P(a \leq x \leq b) = P(a < x \leq b) = P(a \leq x < b) = P(a < x < b)$$

Probabilities for continuous random variables are often calculated using cumulative areas. A cumulative area is all of the area under the density curve to the left of a particular value. Figure 7.8 illustrates the cumulative area to the left of 0.5, which represents $P(x < 0.5)$.

Figure 7.8
A cumulative area under a density curve.



The probability that x is in any particular interval, $P(a < x < b)$, can be expressed as the difference between two cumulative areas.

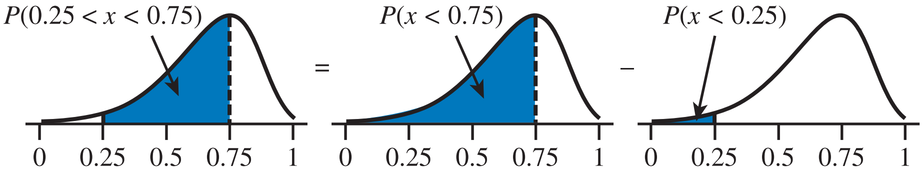
The probability that a continuous random variable x lies between a lower limit a and an upper limit b is

$$\begin{aligned} P(a < x < b) &= (\text{cumulative area to the left of } b) - (\text{cumulative area to the left of } a) \\ &= P(x < b) - P(x < a) \end{aligned}$$

For continuous random variables, the area above an interval of values does not depend on whether either endpoint is included, so this is also the probability of $P(a \leq x \leq b)$, $P(a < x \leq b)$, and $P(a \leq x < b)$.

This property is illustrated in Figure 7.9 for the case of $a = 0.25$ and $b = 0.75$. This result will be used in Section 7.6 when calculating probabilities for random variables that have a normal distribution.

Figure 7.9
Calculation of $P(a < x < b)$ using cumulative areas.

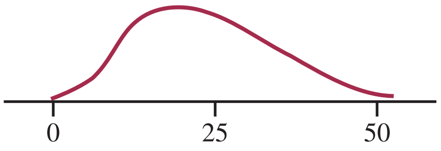


Copyright 2025 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s). Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

For some continuous distributions, cumulative areas can be calculated using methods from the branch of mathematics called calculus. However, because knowledge of calculus is not assumed, technology or on tables that have been constructed for the commonly encountered continuous probability distributions are used. Many graphing calculators and statistical software packages will compute areas for widely used continuous probability distributions.

Exercises 7.22–7.32

7.22 Let x denote the lifetime (in thousands of hours) of a certain type of fan used in diesel engines. The density curve of x is as pictured:



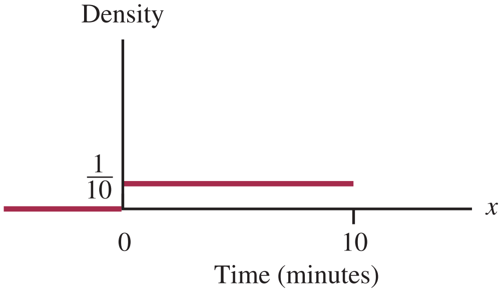
Shade the area under the curve corresponding to each of the following probabilities (draw a new curve for each part). (Hint: See Example 7.8.)

a. $P(10 < x < 25)$
b. $P(10 \leq x \leq 25)$
c. $P(x < 30)$

7.23 Using the density curve for fan lifetime given in the previous exercise, shade the area under the curve corresponding to each of the following probabilities (draw a new curve for each part).

a. The probability that the lifetime is at least 25,000 hours
b. The probability that the lifetime exceeds 25,000 hours

7.24 A particular professor never dismisses class early. Let x denote the amount of time past the class end time (in minutes) that passes before the professor dismisses class. Suppose that x has a uniform distribution on the interval from 0 to 10 minutes. The density curve is shown in the following figure:



a. What is the probability that at most 5 minutes pass before dismissal? (Hint: See Example 7.8.)
b. What is the probability that between 3 and 5 minutes pass before dismissal?

7.25 Refer to the probability distribution given in the previous exercise. Put the following probabilities in order, from smallest to largest:

$$P(2 < x < 3), P(2 \leq x \leq 3), P(x < 2), P(x > 7)$$

Explain your reasoning.

7.26 The article **"Probabilistic Risk Assessment of Infrastructure Networks Subjected to Hurricanes"** ***(12th International Conference on Applications of Statistics and Probability in Civil Engineering*, 2015)** suggests a uniform distribution as a model for the actual landfall position of the eye of a hurricane. Consider the random variable x = distance of actual landfall from predicted landfall. Suppose that a uniform distribution on the interval from 0 km to 400 km is a reasonable model for x.

a. Draw the density curve for x.
b. What is the height of the density curve?

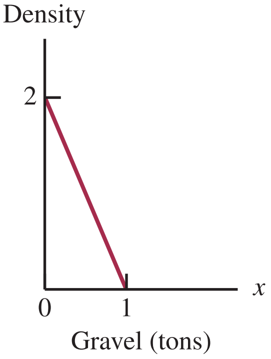
7.27 Use the density curve of x = distance of actual landfall from predicted landfall from the previous exercise to answer the following questions.

a. What is the probability that x is at most 100?
b. What is the probability that x is between 200 and 300?
c. What is the probability that x is between 50 and 150?
d. Explain why the probabilities calculated in Parts (b) and (c) are equal.

7.28 Let x denote the amount of gravel sold (in tons) during a randomly selected week at a particular sales facility. Suppose that the density curve has height $f(x)$ above the value x, where

$$f(x) = \begin{cases} 2(1 - x) & 0 \leq x \leq 1 \\ 0 & \text{otherwise} \end{cases}$$

The density curve (the graph of $f(x)$) is shown in the following figure:



Use the fact that the

$$\text{Area of a triangle} = \frac{1}{2}(\text{Base})(\text{Height})$$

Copyright 2025 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s). Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

to calculate each of the following probabilities. (Hint: Drawing a picture and shading the appropriate area will help.)

a. $P\left(x < \frac{1}{2}\right)$

b. $P\left(x \le \frac{1}{2}\right)$

c. $P\left(x < \frac{1}{4}\right)$

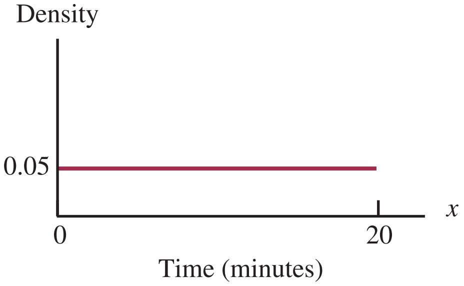
d. $P\left(\frac{1}{4} < x < \frac{1}{2}\right)$ (Hint: Use the results of Parts (a)–(c).)

7.29 Use the density curve for x = amount of gravel sold given in the previous exercise to determine each of the following probabilities.

a. The probability that gravel sold exceeds $\frac{1}{2}$ ton

b. The probability that gravel sold is at least $\frac{1}{4}$ ton

7.30 Let x be the amount of time (in minutes) that a particular San Francisco commuter must wait for a train. Suppose that the density curve is as pictured (a uniform distribution):



a. What is the probability that x is less than 10 minutes? more than 15 minutes?

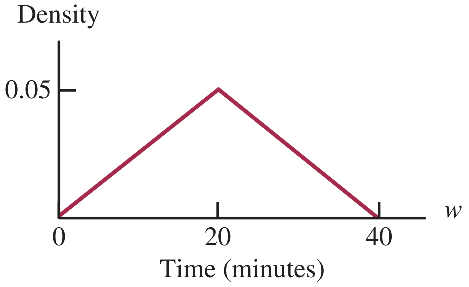
b. What is the probability that x is between 7 and 12 minutes?

c. Find the value c for which $P(x < c) = 0.9$.

7.31 Referring to the previous exercise, let x and y be waiting times on two independently selected days. Define a new random variable w by

$$w = x + y$$

the sum of the two waiting times. The set of possible values for w is the interval from 0 to 40 (because both x and y can range from 0 to 20). It can be shown that the density curve of w is as pictured (this curve is called a triangular distribution, for obvious reasons!):



a. Verify that the total area under the density curve is equal to 1.
(Hint: Area of a triangle = $\frac{1}{2}$(Base)(Height).)

b. What is the probability that w is less than 20?

c. What is the probability that w is less than 10?

d. What is the probability that w is greater than 30?

7.32 The density curve for the random variable w (the sum of two wait times) is given in the previous exercise. What is the probability that w is between 10 and 30? (Hint: It might be easier first to find the probability that w is not between 10 and 30.)

Section 7.4 Mean and Standard Deviation of a Random Variable

When you study the behavior of a random variable x, such as the number of insurance claims made by a homeowner (a discrete variable) or the birth weight of a baby (a continuous variable), you can learn something about how its values are distributed along the measurement scale. The sample mean $\bar{x}$ and sample standard deviation s summarize center and spread for the values in a sample. Similarly, the mean value and standard deviation of a random variable describe where the variable's probability distribution is centered and the extent to which it spreads out about the center.

The **mean value of a random variable x**, denoted by μ_x, describes where the probability distribution of x is centered.

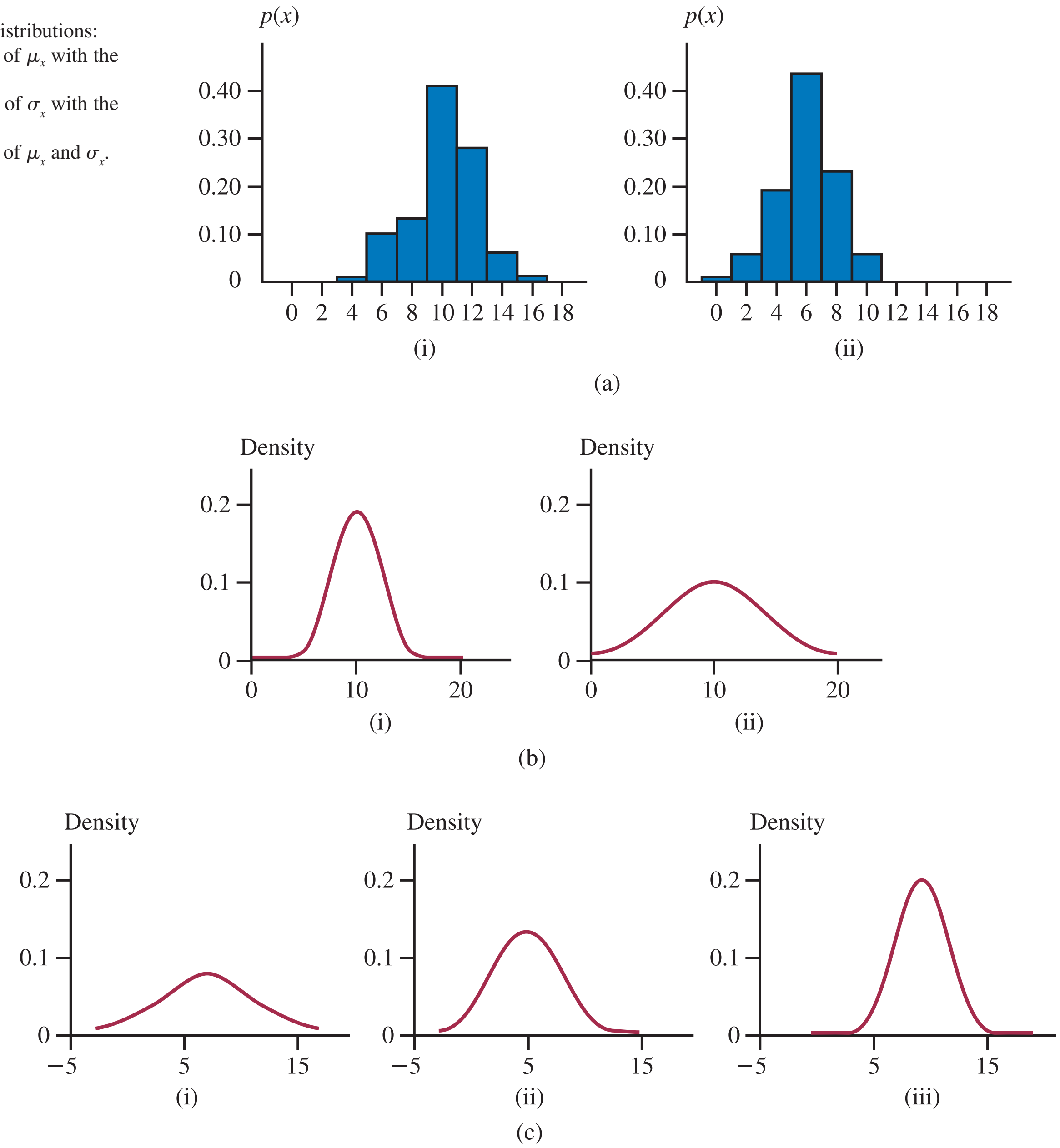
The **standard deviation of a random variable x**, denoted by σ_x, describes variability in the probability distribution. When the value of σ_x is small, observed values of x will tend to be close to the mean value (little variability). When the value of σ_x is large, there will be more variability in observed x values.

Copyright 2025 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s). Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

Consider the probability distributions in Figure 7.10.

1. Figure 7.10(a) shows two discrete probability distributions with the same standard deviation (variability) but different means (center). One distribution has a mean of $\mu_x = 6$ and the other has $\mu_x = 10$. Which has the mean of 6 and which has the mean of 10?
2. Figure 7.10(b) shows two continuous probability distributions that have the same mean but different standard deviations. Which distribution—(i) or (ii)—has the greater standard deviation?
3. Figure 7.10(c) shows three continuous distributions with different means and standard deviations. Which of the three distributions has the greatest mean? Which has a mean of about 5? Which distribution has the smallest standard deviation?

Figure 7.10
Some probability distributions:
(a) different values of μ_x with the same value of σ_x;
(b) different values of σ_x with the same value of μ_x;
(c) different values of μ_x and σ_x.



The answers to these questions are the following: Figure 7.10(a)(ii) has a mean of 6, and Figure 7.10(a)(i) has a mean of 10; Figure 7.10(b)(ii) has the greater standard deviation; Figure 7.10(c)(iii) has the greatest mean, Figure 7.10(c)(ii) has a mean of about 5, and Figure 7.10(c)(iii) has the smallest standard deviation.

It is customary to use the terms mean of the random variable x and mean of the probability distribution of x interchangeably. Similarly, the standard deviation of the random variable x and the standard deviation of the probability distribution of x refer to the same thing. Although the mean and standard deviation are calculated differently for discrete and continuous random variables, the interpretation is the same in both cases.

Mean Value of a Discrete Random Variable

Consider a chance experiment consisting of randomly selecting an automobile licensed in Pennsylvania. Define the discrete random variable x as

x = the number of low-beam headlights on the selected car that need adjustment

Possible x values are 0, 1, and 2, and the probability distribution of x might be as follows:

x value	0	1	2
Probability	0.5	0.3	0.2

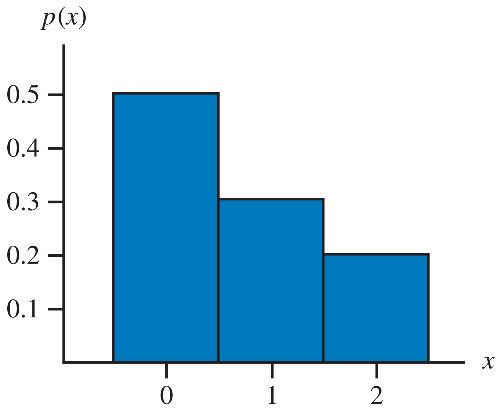
The corresponding probability histogram is given in Figure 7.11. In a sample of 100 cars licensed in Pennsylvania, the sample relative frequencies might differ somewhat from the given probabilities (which are the limiting relative frequencies). For example, you might see:

x value	0	1	2
Frequency	46	33	21

The sample mean value of x for these 100 observations is the sum of 46 zeros, 33 ones, and 21 twos, all divided by 100:

$$\begin{aligned}\bar{x} &= \frac{(46)(0) + (33)(1) + (21)(2)}{100}\\ &= \left(\frac{46}{100}\right)(0) + \left(\frac{33}{100}\right)(1) + \left(\frac{21}{100}\right)(2)\\ &= (\text{rel. freq. of } 0)(0) + (\text{rel. freq. of } 1)(1) + (\text{rel. freq. of } 2)(2)\\ &= 0.75\end{aligned}$$

Figure 7.11 Probability histogram for the distribution of the number of headlights needing adjustment.



As the sample size increases, each relative frequency will approach the corresponding probability. In a very long sequence of observations, the value of $\bar{x}$ will approach

$$\begin{aligned}\text{mean value of } x &= P(x = 0)(0) + P(x = 1)(1) + P(x = 2)(2)\\ &= (0.5)(0) + (0.3)(1) + (0.2)(2)\\ &= 0.70\end{aligned}$$

Notice that the expression for $\bar{x}$ is a weighted average of possible x values. The weight of each value is the observed relative frequency. Similarly, the mean value of the discrete random variable x is a weighted average, but with weights that are the probabilities from the probability distribution.

Definition

Mean value of a discrete random variable: The mean value of a discrete random variable x, denoted by μ_x, is calculated by first multiplying each possible x value by the probability of observing that value and then adding the resulting quantities:

$$\mu_x = \sum_{\substack{\text{all possible} \\ x \text{ values}}} x \cdot p(x)$$

The term **expected value** is sometimes used in place of mean value, and $\boldsymbol{E(x)}$ is another way to denote μ_x.

Example 7.9 Exam Attempts

Understand the context ➤ Suppose that students taking an online course must take an exam and they are allowed to retake the exam up to three additional times in order to improve their scores. Consider the chance experiment of selecting a student at random. Suppose that the random variable x defined as

$x =$ the number of times the selected student took the exam

Suppose the probability distribution of x is as follows:

x	1	2	3	4
$p(x)$	0.10	0.20	0.30	0.40

Do the work ➤ Then x has mean value

$$\begin{aligned}\mu_x &= \sum_{x=1,2,3,4} x \cdot p(x) \\ &= (1)p(1) + (2)p(2) + (3)p(3) + (4)p(4) \\ &= (1)(0.10) + (2)(0.20) + (3)(0.30) + (4)(0.40) \\ &= 0.10 + 0.40 + 0.90 + 1.60 \\ &= 3.00\end{aligned}$$

Interpret the results ➤ This means that for students in the online course, the mean number of times the exam is taken is 3.

Example 7.10 Apgar Scores

Understand the context ➤ At 1 minute after birth and again at 5 minutes, each newborn child is given a numerical rating called an Apgar score. Possible values of this score are 0, 1, 2, . . . , 9, 10. A child's score is determined by five factors: muscle tone, skin color, respiratory effort, strength of heartbeat, and reflex, with a high score indicating a healthy infant.

Consider the random variable x defined as

$x =$ the Apgar score (at 1 minute) of a randomly selected newborn infant at a particular hospital

Suppose that x has the following probability distribution:

x	0	1	2	3	4	5	6	7	8	9	10
$p(x)$	0.002	0.001	0.002	0.005	0.02	0.04	0.17	0.38	0.25	0.12	0.01

Do the work ➤ The mean value of x is

$$\begin{aligned}\mu_x &= (0)p(0) + (1)p(1) + \cdots + (9)p(9) + (10)p(10) \\ &= (0)(0.002) + (1)(0.001) + \cdots + (9)(0.12) + (10)(0.01) \\ &= 7.16\end{aligned}$$

Interpret the results ➤ The mean Apgar score for babies born at this hospital is 7.16. The mean Apgar score for a *sample* of newborn children born at this hospital might be $\bar{x} = 7.05$, $\bar{x} = 8.30$, or any one of a number of other possible values between 0 and 10. However, as child after child is born

Copyright 2025 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s). Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

and rated, the mean Apgar score will approach the value 7.16, which can be interpreted as the mean Apgar score for the population of all babies born at this hospital.

Standard Deviation of a Discrete Random Variable

The mean value μ_x provides only a partial summary of a probability distribution. Two different distributions can have the same value of μ_x, but there may be more variability in a long sequence of observations from one distribution than in a long sequence of observations from the other distribution.

Example 7.11 Glass Panels

Understand the context ➤

Flat screen TVs require high quality glass with very few flaws. Suppose a television manufacturer receives glass panels from two different suppliers. Let x denote the number of flaws in a randomly selected glass panel from the first supplier and y denote the number of flaws in a randomly selected glass panel from the second supplier. Suppose that the probability distributions for x and y are as follows:

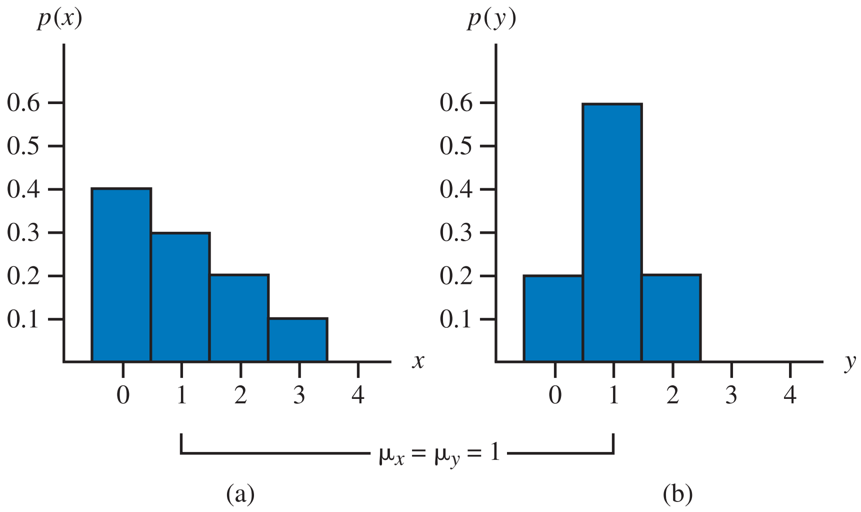
x	0	1	2	3	4
$p(x)$	0.4	0.3	0.2	0.1	0

y	0	1	2	3	4
$p(y)$	0.2	0.6	0.2	0	0

Probability histograms for x and y are given in Figure 7.12.

It is easy to verify that the mean values of both x and y are 1, so for either supplier the long-run average number of flaws per panel is 1. However, the two probability histograms show that the probability distribution for the second supplier is concentrated closer to the mean value than is the case for the first supplier.

Figure 7.12
Probability distribution for the number of flaws in a glass panel in Example 7.11:
(a) Supplier 1;
(b) Supplier 2.



Interpret the results ➤

The greater spread of the first distribution implies that there will be more variability in a long sequence of observed x values than in a long sequence of observed y values. For example, the y sequence will contain no 3's, but in the long run, 10% of the observed x values will be 3.

Calculation of the variance and standard deviation of a random variable x involves squared deviations from the mean. A value far from the mean results in a large squared deviation. However, such a value does not contribute substantially to variability in x if the probability associated with that value is small. For example, if $\mu_x = 1$ and $x = 25$ is a possible value, then the squared deviation is $(25 - 1)^2 = 576$. However, if $P(x = 25) = 0.000001$, the value 25 will hardly ever be observed, so it won't contribute much to variability in a long sequence of observations. This is why each squared deviation is multiplied by the probability associated with the value to obtain a measure of variability.

Definitions

Variance of a discrete random variable: The variance of a discrete random variable x, denoted by σ_x^2, is calculated by

1. subtracting the mean from each possible x value to obtain the deviations
2. squaring each deviation
3. multiplying each squared deviation by the probability of the corresponding x value
4. adding these quantities

$$\sigma_x^2 = \sum_{\substack{\text{all possible} \\ x \text{ values}}} (x - \mu_x)^2 p(x)$$

Standard deviation of a discrete random variable: The standard deviation of a discrete random variable x, denoted by σ_x, is the square root of the variance.

Example 7.12 Glass Panels Revisited

For x = number of flaws in a glass panel from the first supplier in Example 7.11,

$$\begin{aligned}\sigma_x^2 &= (0-1)^2p(0) + (1-1)^2p(1) + (2-1)^2p(2) + (3-1)^2p(3) \\ &= (1)(0.4) + (0)(0.3) + (1)(0.2) + (4)(0.1) \\ &= 1.0\end{aligned}$$

The standard deviation of x is then $\sigma_x = \sqrt{\sigma_x^2} = \sqrt{1.0} = 1.0$.

For y = the number of flaws in a glass panel from the second supplier,

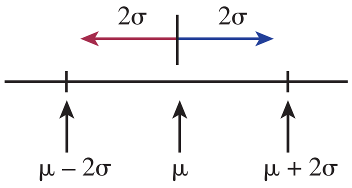
$$\sigma_y^2 = (0-1)^2(0.2) + (1-1)^2(0.6) + (2-1)^2(0.2) = 0.4$$

Then $\sigma_y = \sqrt{0.4} = 0.632$.

The fact that σ_x is greater than σ_y confirms the impression about the variability in x and y conveyed by the probability distributions shown in Figure 7.12.

Example 7.13 More on Apgar Scores

Figure 7.13
Values within 2 standard deviations of the mean.



Reconsider the distribution of Apgar scores for children born at a certain hospital, introduced in Example 7.10. What is the probability that a randomly selected child's score will be within 2 standard deviations of the mean score? As Figure 7.13 shows, values of x within 2 standard deviations of the mean are those for which

$$\mu - 2\sigma < x < \mu + 2\sigma$$

From Example 7.10 you already know $\mu_x = 7.16$. The variance is

$$\begin{aligned}\sigma^2 &= \Sigma(x-\mu)^2p(x) = \Sigma(x - 7.16)^2p(x) \\ &= (0-7.16)^2(0.002) + (1-7.16)^2(0.001) + \cdots + (10-7.16)^2(0.01) \\ &= 1.5684\end{aligned}$$

and the standard deviation is

$$\sigma = \sqrt{1.5684} = 1.25$$

Copyright 2025 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s). Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

This gives (using the probabilities given in Example 7.10)

$$\begin{aligned} P(\mu - 2\sigma < x < \mu + 2\sigma) &= P(7.16 - 2.50 < x < 7.16 + 2.50) \\ &= P(4.66 < x < 9.66) \\ &= p(5) + \cdots + p(9) \\ &= 0.96 \end{aligned}$$

Mean and Standard Deviation When *x* Is Continuous

For continuous probability distributions, μ_x and σ_x are defined and calculated using methods from calculus. You don't need to worry about calculating these values. What is important is knowing that μ_x and σ_x are interpreted in the same way as they are in the discrete case. The mean value μ_x locates the center of the continuous distribution. It gives the approximate long-run average of many observed x values. The standard deviation σ_x measures the extent that the continuous distribution (density curve) spreads out around μ_x. It provides information about the variability that can be expected in a long sequence of observed x values.

Example 7.14 Body Mass Index

Understand the context ➤

Body mass index (BMI) is a continuous variable calculated from height and weight that is measured in kilograms/meter^2. The authors of the paper **"Concordance of Self-Report and Measured Height and Weight of College Students" (*Journal of Nutrition Education and Behavior* [2015]: 94–98)** collected data on BMI from a large number of students at eight different colleges. For these students, the authors reported that the continuous random variables

x = BMI based on self-reported height and weight

and

y = BMI based on measured height and weight

have the following means and standard deviations:

$$\begin{array}{ll} \mu_x = 24.2 & \mu_y = 24.5 \\ \sigma_x = 2.6 & \sigma_y = 3.9 \end{array}$$

Interpret the results ➤

This suggests that if you were to observe self-reported and measured BMI for a large number of these college students, you would not see much difference in the means of the observations for these variables. But because the standard deviation of y = measured BMI is larger than the standard deviation of x = self-reported BMI, there would be more variability in the values of the measured BMI than the BMI that was based on self-reported height and weight.

Mean and Variance of Linear Functions and Linear Combinations

You have seen how the mean and standard deviation of one or more random variables provide useful information about the variables' long-run behavior. Sometimes you are also interested in the behavior of some function of random variables.

For example, consider the chance experiment in which a customer of a propane gas company is randomly selected. Suppose that the mean and standard deviation of the random variable

x = number of gallons required to fill a customer's propane tank

are known to be 318 gallons and 42 gallons, respectively. The company is considering two different pricing models:

Model 1: \$3 per gallon
Model 2: service charge of \$50 + \$2.80 per gallon

The company is interested in the variable

y = amount billed

For each of the two pricing models, y can be expressed as a function of the random variable x:

$$\text{Model 1: } y_{model\ 1} = 3x$$
$$\text{Model 2: } y_{model\ 2} = 50 + 2.8x$$

Both of these equations are examples of a linear function of the random variable x. The mean and standard deviation of a linear function of x can be calculated from the mean and standard deviation of x, as described in the following box.

The Mean, Variance, and Standard Deviation of a Linear Function

If x is a random variable with mean μ_x and variance σ_x^2 and a and b represent numbers, the random variable y defined by

$$y = a + bx$$

is called a **linear function of the random variable *x*.**

The mean of $y = a + bx$ is

$$\mu_y = \mu_{a+bx} = a + b\mu_x$$

The variance of y is

$$\sigma_y^2 = \sigma_{a+bx}^2 = b^2\sigma_x^2$$

It follows that the standard deviation of y is

$$\sigma_y = \sigma_{a+bx} = |b|\sigma_x$$

You can use the results in the preceding box to calculate the mean and standard deviation of the billing amount variable for the propane gas example, as follows:

For Model 1:

$$\mu_{model\ 1} = \mu_{3x} = 3\mu_x = 3(318) = 954$$
$$\sigma_{model\ 1}^2 = \sigma_{3x}^2 = 3^2\sigma_x^2 = 9(42)^2 = 15{,}876$$
$$\sigma_{model\ 1} = \sqrt{15{,}876} = 126, \text{ which is equal to } 3(42)$$

For Model 2:

$$\mu_{model\ 2} = \mu_{50+2.8x} = 50 + 2.8\mu_x = 50 + 2.8(318) = 940.40$$
$$\sigma_{model\ 2}^2 = \sigma_{50+2.8x}^2 = 2.8^2\sigma_x^2 = (2.8)^2(42)^2 = 13{,}829.76$$
$$\sigma_{model\ 2} = \sqrt{13{,}829.76} = 117.60, \text{ which is equal to } 2.8(42)$$

The mean billing amount for Model 1 is a bit greater than for Model 2, as is the variability in billing amounts. Billing amounts based on Model 2 will be slightly more consistent from bill to bill.

Linear Combinations

Now consider a different type of problem. Suppose that you have three tasks that you plan to complete on the way home from school: stop at the library to return an overdue book for which you must pay a fine, deposit your most recent paycheck at the bank, and stop by the office supply store to buy paper for your printer. Define the following variables:

x_1 = time required to return book and pay fine
x_2 = time required to deposit paycheck
x_3 = time required to buy printer paper

Copyright 2025 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s). Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

You can then define a new variable, y, to represent the total amount of time to complete these tasks:

$$y = x_1 + x_2 + x_3$$

Defined in this way, y is an example of a linear combination of random variables.

If $x_1, x_2, \ldots, x_n$ are random variables and $a_1, a_2, \ldots, a_n$ represent numbers, the random variable y defined as

$$y = a_1x_1 + a_2x_2 + \cdots + a_nx_n$$

is a **linear combination of random variables**.

For example, $y = 10x_1 - 5x_2 + 8x_3$ is a linear combination of x_1, x_2, and x_3 with $a_1 = 10$, $a_2 = -5$ and $a_3 = 8$.

It is easy to calculate the mean of a linear combination of random variables if the individual variable means are known. The variance and standard deviation of a linear combination of random variables are also easily calculated *if the random variables are independent*. Two random variables x_i and x_j are independent if any event defined solely by x_i is independent of any event defined solely by x_j. When the x_i's are not independent, calculation of the variance and standard deviation of a linear combination of random variables is more complicated, and this case is not considered here.

Mean, Variance, and Standard Deviation for Linear Combinations

Suppose $x_1, x_2, \ldots, x_n$ are random variables with means $\mu_1, \mu_2, \ldots, \mu_n$ and variances $\sigma_1^2, \sigma_2^2, \ldots, \sigma_n^2$, respectively. Define the random variable y as

$$y = a_1\,x_1 + a_2\,x_2 + \cdots + a_n\,x_n$$

Then

1. $\mu_y = \mu_{a_1x_1+a_2x_2+\cdots+a_nx_n} = a_1\mu_1 + a_2\mu_2 + \cdots + a_n\mu_n$

   This result is true regardless of whether the x_i's are independent.

2. When $x_1, x_2, \ldots, x_n$ are *independent* random variables,

   $$\sigma_y^2 = \sigma^2_{a_1x_1+a_2x_2+\cdots+a_nx_n} = a_1^2\sigma_1^2 + a_2^2\sigma_2^2 + \cdots + a_n^2\sigma_n^2$$

   $$\sigma_y = \sigma_{a_1x_1+a_2x_2+\cdots+a_nx_n} = \sqrt{a_1^2\sigma_1^2 + a_2^2\sigma_2^2 + \cdots + a_n^2\sigma_n^2}$$

   These formulas are appropriate only when the x_i's are independent.

Examples 7.15–7.17 illustrate the use of these rules.

Example 7.15 Freeway Traffic

Understand the context ➤

Three different roads feed into a particular freeway entrance. Suppose that during a fixed time period, the number of cars coming from each road onto the freeway is a random variable with mean values as follows:

Road	1	2	3
Mean	800	1000	600

With x_i representing the number of cars entering from road i, you can define the total number of cars entering the freeway y as

$$y = x_1 + x_2 + x_3.$$

Do the work ➤ The mean value of y is

$$\begin{aligned}\mu_y &= \mu_{x_1+x_2+x_3}\\ &= \mu_{x_1} + \mu_{x_2} + \mu_{x_3}\\ &= 800 + 1000 + 600\\ &= 2400\end{aligned}$$

Example 7.16 Combining Exam Subscores

Understand the context ➤ A nationwide standardized exam consists of a multiple-choice section and a free response section. Suppose that for each section, the mean and standard deviation are reported to be as shown in the following table:

	Mean	**Standard Deviation**
Multiple Choice	38	6
Free Response	30	7

Define x_1 as the multiple-choice score and x_2 as the free-response score for a student selected at random from those taking this exam. Also of interest is the variable y = overall exam score.

Suppose that the free-response score is given twice the weight of the multiple-choice score in determining the overall exam score. Then the overall score is calculated as

$$y = x_1 + 2x_2$$

Do the work ➤ What are the mean and standard deviation of y?

Because $y = x_1 + 2x_2$ is a linear combination of x_1 and x_2, the mean of y is

$$\begin{aligned}\mu_y &= \mu_{x_1+2x_2}\\ &= \mu_{x_1} + 2\mu_{x_2}\\ &= 38 + 2(30)\\ &= 98\end{aligned}$$

Interpret the results ➤ This means that the average overall score for students taking this exam is 98. What about the variance and standard deviation of y? To use the formulas in the preceding box, x_1 and x_2 must be independent. It is unlikely that the value of x_1 (a student's multiple-choice score) would be unrelated to the value of x_2 (the same student's free-response score), because it seems likely that students who score well on one section of the exam will also tend to score well on the other section. Therefore, it would not be appropriate to use the formulas given previously to calculate the variance and standard deviation.

Example 7.17 Baggage Weights

Understand the context ➤ A commuter airline flies small planes between San Luis Obispo, California, and San Francisco. For small planes, baggage weight is a concern, especially on foggy mornings, because the weight of the plane has an effect on how quickly the plane can ascend. Suppose that it is known that the random variable x = weight (in pounds) of baggage checked by a randomly selected passenger has a mean of 42 pounds and standard deviation of 16.

Consider a flight with 10 passengers, all traveling alone. If you use x_i to denote the baggage weight for passenger i (for i ranging from 1 to 10), the total weight of checked baggage, y, is then

$$y = x_1 + x_2 + \cdots + x_{10}$$

Copyright 2025 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s). Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

Do the work ➤ Notice that y is a linear combination of the x_i's. The mean value of y is

$$\begin{aligned}\mu_y &= \mu_{x_1} + \mu_{x_2} + \cdots + \mu_{x_{10}} \\ &= 42 + 42 + \cdots + 42 \\ &= 420\end{aligned}$$

Since the 10 passengers are all traveling alone, it is reasonable to think that the 10 baggage weights are unrelated and that the x_i's are independent. (This might not be a reasonable assumption if the 10 passengers were not traveling alone.) Then the variance of y is

$$\begin{aligned}\sigma^2_y &= \sigma^2_{x_1} + \sigma^2_{x_2} + \cdots + \sigma^2_{x_{10}} \\ &= 16^2 + 16^2 + \cdots + 16^2 \\ &= 2650\end{aligned}$$

and the standard deviation of y is

$$\sigma_y = \sqrt{2650} = 50.6$$

Interpret the results ➤ This means that the mean total weight of checked baggage is 420 pounds for flights with 10 passengers traveling alone. The large standard deviation of 50.6 pounds indicates that there will be a lot of variability in total checked baggage weight from flight to flight.

One Last Note on Linear Functions and Linear Combinations

In Example 7.17, the random variable of interest was x = weight of baggage checked by a randomly selected airline passenger. However, when a flight with multiple passengers was considered, subscripts were added to create variables like

x_1 = baggage weight for passenger 1

and

x_2 = baggage weight for passenger 2

It is important to note that the sum of the baggage weights for two different passengers does not result in the same value as doubling the baggage weight of a single passenger. *The linear combination* $x_1 + x_2$ *is different from the linear function* $2x$.

Although the mean of $x_1 + x_2$ and the mean of $2x$ are the same, the variances and standard deviations are different. For example, for the baggage weight distribution described in Example 7.17 (mean of 42 and standard deviation of 16), the mean of $2x$ is

$$\mu_{2x} = 2\mu_x = 2(42) = 84$$

and the mean of $x_1 + x_2$ is

$$\mu_{x_1+x_2} = \mu_{x_1} + \mu_{x_2} = 42 + 42 = 84$$

However, the variance of $2x$ is

$$\sigma^2_{2x} = (2)^2\sigma^2_x = 4(16^2) = 1024$$

and the variance of $x_1 + x_2$ is

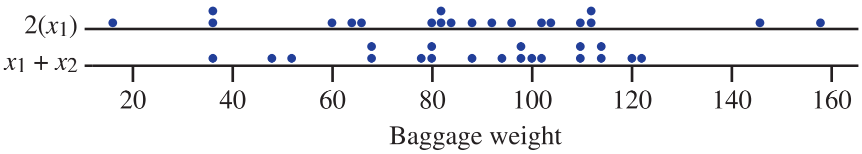
$$\sigma^2_{x_1+x_2} = \sigma^2_{x_1} + \sigma_{2x_2} = 16^2 + 16^2 = 256 + 256 = 512$$

This means that observations of $2x$ for randomly selected passengers will show more variability than observations of $x_1 + x_2$ for two randomly selected passengers. This may seem counterintuitive, but it is true! For example, each row of the first two columns of the accompanying table show simulated baggage weights for two passengers (an x_1 and an x_2 value) that were selected at random from a probability distribution with mean 42 and standard deviation 16. The third column of the table contains the value of $2x_1$ (two times the passenger 1 baggage weight) and the last column contains the values of $x_1 + x_2$ (the sum of the passenger 1 baggage weight and the passenger 2 baggage weight).

x_1	x_2	$2x_1$	$x_1 + x_2$	x_1	x_2	$2x_1$	$x_1 + x_2$
56	41	112	97	44	23	88	67
30	5	60	35	48	54	96	102
32	15	64	47	52	69	104	121
55	45	110	100	56	63	112	119
40	54	80	94	79	35	158	114
41	68	82	109	18	34	36	52
42	35	84	77	8	71	16	79
33	47	66	80	51	58	102	109
46	52	92	98	18	49	36	67
73	40	146	113	41	46	82	87

Figure 7.14 shows dotplots of the 20 values of $2x_1$ and the 20 values of $x_1 + x_2$. You can see that there is less variability in the $x_1 + x_2$ values than in the values of $2x_1$.

Figure 7.14
Dotplots of $2x_1$ and $x_1 + x_2$.



Exercises 7.33–7.50

7.33 Consider selecting a household in the United States at random. Define the random variable x to be

x = number of individuals living in the selected household

Based on information from the U.S. Census Bureau in 2021, the probability distribution of x is given below. Households with more than seven individuals were included in the probability value for 7.

x	1	2	3	4	5	6	7
$p(x)$	0.2845	0.3503	0.1503	0.1239	0.0583	0.0203	0.0124

a. Calculate the mean and standard deviation of x. (Hint: See Examples 7.9 and 7.12.)

b. The mean household size calculated in Part (a) treated all households with seven or more individuals as if they had only seven individuals. How do you think the mean calculated in Part (a) compares to the actual mean household size in 2021?

c. The U.S. Census Bureau reported a mean household size for 2021 of 2.54. How does this value compare to the mean that you calculated in Part (a)? Does this support your answer in Part (b)?

7.34 Suppose the probability distribution of x, the number of defective tires on a randomly selected car checked at an inspection station, is given in the following table:

x	0	1	2	3	4
$p(x)$	0.54	0.16	0.06	0.04	0.20

a. Calculate the mean value of x.

b. What is the probability that x exceeds its mean value?

7.35 Consider the following probability distribution for y = the number of broken eggs in a carton:

y	0	1	2	3	4
$p(y)$	0.65	0.20	0.10	0.04	0.01

a. Calculate and interpret μ_y.

b. In the long run, for what percentage of cartons is the number of broken eggs less than μ_y? Does this surprise you?

c. Why doesn't $\mu_y = (0 + 1 + 2 + 3 + 4)/5 = 2.0$? Explain.

7.36 Referring to the previous exercise, use the result of Part (a) along with the fact that a carton contains 12 eggs to determine the mean value of z = the number of unbroken eggs. (Hint: z can be written as a linear function of y; see Example 7.15.)

Copyright 2025 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s). Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

7.37 Exercise 7.8 gave the following probability distribution for x = the number of courses for which a randomly selected student at a certain university is registered:

x	1	2	3	4	5	6	7
$p(x)$	0.02	0.03	0.09	0.25	0.40	0.16	0.05

For this probability distribution, $\mu = 4.66$ and $\sigma = 1.20$.

a. Because $\mu - \sigma = 3.46$, the x values 1, 2, and 3 are more than 1 standard deviation below the mean. What is the probability that x is more than 1 standard deviation below its mean? (Hint: See Example 7.13.)

b. What x values are more than 2 standard deviations away from the mean value (either less than $\mu - 2\sigma$ or greater than $\mu + 2\sigma$)?

c. What is the probability that x is more than 2 standard deviations away from its mean value?

7.38 Example 7.11 gave the probability distributions of x = number of flaws in a randomly selected glass panel for two suppliers of glass used to make flat screen TVs. If the manufacturer wanted to select a single supplier for glass panels, which of these two suppliers would you recommend? Justify your choice based on consideration of both center and variability. (Hint: See Example 7.11.)

7.39 Consider a large ferry that can accommodate cars and buses. The toll for cars is \$3, and the toll for buses is \$10. Let x and y denote the number of cars and buses, respectively, carried on a single trip. Cars and buses are accommodated on different levels of the ferry, so the number of buses accommodated on any trip is independent of the number of cars on the trip. Suppose that x and y have the following probability distributions:

x	0	1	2	3	4	5
$p(x)$	0.05	0.10	0.25	0.30	0.20	0.10

y	0	1	2
$p(y)$	0.50	0.30	0.20

a. Calculate the mean and standard deviation of x.

b. Calculate the mean and standard deviation of y.

7.40 Refer to the information given in Exercise 7.39.

a. Calculate the mean and variance of the total amount of money collected in tolls from cars.

b. Calculate the mean and variance of the total amount of money collected in tolls from buses.

7.41 Refer to the information given in Exercise 7.39.

a. Calculate the mean and variance of z = total number of vehicles (cars and buses) on the ferry.

b. Calculate the mean and variance of w = total amount of money collected in tolls.

7.42 Suppose that for a particular real estate agent, the probability distribution of x = the number of houses sold in 1 month is given by the following table:

x	1	2	3	4	5	6	7	8
$p(x)$	0.05	0.10	0.12	0.30	0.30	0.11	0.01	0.01

a. Find the mean value of x (the mean number of houses sold).

b. Find the variance and standard deviation of x. How would you interpret these values?

c. What is the probability that the number of houses sold is within 1 standard deviation of its mean value?

d. What is the probability that the number of houses sold is more than 2 standard deviations from the mean? (Hint: See Example 7.13.)

7.43 A local television station sells 15-second, 30-second, and 60-second advertising spots. Let x denote the length of a randomly selected commercial appearing on this station. Suppose that the probability distribution of x is given by the following table:

x	15	30	60
$p(x)$	0.1	0.3	0.6

a. Find the mean length for commercials appearing on this station.

b. If a 15-second spot sells for \$500, a 30-second spot for \$800, and a 60-second spot for \$1000, find the average amount paid for commercials appearing on this station. (Hint: Consider a new variable, y = cost, and then find the probability distribution and mean value of y.)

7.44 An author has written a book and submitted it to a publisher. The publisher offers to print the book and gives the author the choice between a flat payment of \$10,000 and a royalty plan. Under the royalty plan the author would receive \$1 for each copy of the book sold. The author thinks that the following table gives the probability distribution of the variable x = the number of books sold:

x	1,000	5,000	10,000	20,000
$p(x)$	0.05	0.30	0.40	0.25

Which payment plan should the author choose? Explain your reasoning.

7.45 A grocery store has an express line for customers purchasing five or fewer items. Let x be the number of items purchased by a randomly selected customer using this line. Give examples of two different assignments of probabilities such that the resulting distributions have the same mean but different standard deviations.

7.46 An appliance dealer sells three different models of upright freezers having 13.5, 15.9, and 19.1 cubic feet of storage space. Let x = the amount of storage space purchased by a customer who buys a freezer. Suppose that x has the following probability distribution:

x	13.5	15.9	19.1
$p(x)$	0.2	0.5	0.3

a. Calculate the mean and standard deviation of x.

b. Suppose the price of the freezer depends on the size of the storage space, x. If $Price = 25x - 8.5$, what is the mean price paid by customers who buy freezers?

c. What is the standard deviation of the price paid?

7.47 To assemble a piece of furniture, a wood peg must be inserted into a predrilled hole. Suppose that the diameter of a randomly selected peg is a random variable with mean 0.25 inch and standard deviation 0.006 inch and that the diameter of a randomly selected hole is a random variable with mean 0.253 inch and standard deviation 0.002 inch. Let x_1 = peg diameter, and let x_2 = hole diameter.

a. Why would the random variable y, defined as $y = x_2 - x_1$, be of interest to the furniture manufacturer?

b. What is the mean value of the random variable y?

c. Assuming that x_1 and x_2 are independent, what is the standard deviation of y?

d. Based on your answers to Parts (b) and (c), do you think that finding a peg that is too big to fit in the predrilled hole would be a relatively common or a relatively rare occurrence? Explain.

7.48 A multiple-choice exam consists of 50 questions. Each question has five choices, of which only one is correct. Suppose that the total score on the exam is calculated as

$$y = x_1 - \frac{1}{4}x_2$$

where x_1 = number of correct responses and x_2 = number of incorrect responses. (Calculating a total score by subtracting a term based on the number of incorrect responses is known as a correction for guessing and is designed to discourage test takers from choosing answers at random.)

a. It can be shown that if a totally unprepared student answers all 50 questions by just selecting one of the five answers at random, then $\mu_{x_1} = 10$ and $\mu_{x_2} = 40$. What is the mean value of the total score, y? Does this surprise you? Explain. (Hint: See Example 7.16.)

b. Explain why it is unreasonable to use the formulas given in this section to compute the variance or standard deviation of y.

7.49 A number cube is a six-sided cube with the numbers 1, 2, 3, 4, 5, and 6, one on each side of the cube. Consider a game in which a red number cube and a blue number cube are rolled. Let x_R denote the value showing on the uppermost face of the red number cube, and define x_B similarly for the blue number cube.

a. The probability distribution of x_R is

x_R	1	2	3	4	5	6
$p(x_R)$	1/6	1/6	1/6	1/6	1/6	1/6

Find the mean, variance, and standard deviation of x_R.

b. What are the values of the mean, variance, and standard deviation of x_B? (You should be able to answer this question without doing any additional calculations.)

7.50 Consider the random variables x_R and x_B defined in the previous exercise.

a. Suppose that you are offered a choice of the following two games:

Game 1: Pay $7 to play, and you win y_1 dollars, where $y_1 = x_R + x_B$.

Game 2: Doesn't cost anything to play initially, but you "win" $3y_2$ dollars, where $y_2 = x_R - x_B$. If y_2 is negative, you must pay that amount; if it is positive, you receive that amount.

For Game 1, the net amount won in a game is $w_1 = y_1 - 7 = x_R + x_B - 7$. What are the mean and standard deviation of w_1?

b. For Game 2, the net amount won in a game is $w_2 = 3y_2 = 3(x_R - x_B)$. What are the mean and standard deviation of w_2?

c. Based on your answers to Parts (a) and (b), if you had to play, which game would you choose? Explain your reasoning.

Section 7.5 Binomial and Geometric Distributions

In this section, two discrete probability distributions are introduced: the binomial distribution and the geometric distribution. These distributions arise when a chance experiment consists of making a sequence of observations when there are two possible values for each observation. The process of making a single observation is called a *trial*.

For example, one characteristic of blood type is Rh factor, which can be either positive or negative. You can think of a chance experiment that consists of noting the Rh factor for

Copyright 2025 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s). Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

each of 25 blood donors as a sequence of 25 trials. Each trial consists of observing the Rh factor of a single donor (there are two possible values–positive or negative).

You might also conduct a different chance experiment that consists of observing the Rh factor of blood donors until a donor who is Rh-negative is encountered. This second experiment can also be viewed as a sequence of trials, but the total number of trials in this experiment is not predetermined, as it was in the previous example, where you knew in advance that there would be 25 trials. Experiments of the two types just described are typical of those leading to the binomial distribution and to the geometric distribution.

Binomial Distributions

Students at a particular college can choose to purchase textbooks in either a digital or print format. Suppose that you plan to interview 25 students at this college. For each student, you will record whether they purchase all of their textbooks in a digital format (with possible responses yes or no). What is the chance that at least 15 purchase all of their books in a digital format? What is the chance that between 10 and 15 are purchasing all of their books in a digital format? How many among the 25 can you expect to purchase all of their books in a digital format? These and other similar questions can be answered by studying the **binomial probability distribution**.

The binomial distribution arises when the chance experiment of interest is a **binomial experiment**. Binomial experiments have the properties listed in the following box.

Properties of a Binomial Experiment

A **binomial experiment** consists of a sequence of trials that meet the following conditions:

1. There are a fixed number of trials.
2. Each trial can result in one of only two possible outcomes, labeled success (S) and failure (F).
3. Outcomes of different trials are independent.
4. The probability that a trial results in a success is the same for each trial.

The **binomial random variable *x*** is defined as

x = number of successes observed when a binomial experiment is performed

The probability distribution of x is called the **binomial probability distribution**.

The term *success* here does not necessarily have its usual meaning. Which of the two possible outcomes is labeled "success" is determined by the random variable of interest. For example, if the variable counts the number of students who purchase all their textbooks in digital format, then purchasing all books in a digital format would be labeled a success (because this is what the variable counts). If the number of students who did not purchase all books in a digital format was counted instead, not purchasing all books in a digital format would be labeled a success and purchasing all books in a digital format would be labeled a failure.

One situation in which a binominal probability distribution arises was given in Example 7.5. In that example, you considered x = number among four customers who purchase an electric or hybrid car (rather than a gas-powered car). This is a binomial experiment with four trials, where the purchase of an electric or hybrid car is considered a success and $P(\text{success}) = P(E) = 0.07$. The 16 possible outcomes, along with the associated probabilities, were displayed in Table 7.1.

Consider now the case of five customers, a binomial experiment with five trials. The possible values of

x = number who purchase an electric or hybrid car

are 0, 1, 2, 3, 4, and 5. There are 32 possible outcomes of the binomial experiment, each one a sequence of five successes and failures. Five of these outcomes result in $x = 1$: *SFFFF*, *FSFFF*, *FFSFF*, *FFFSF*, and *FFFFS*.

Because the trials are independent, the first of these outcomes has probability

$$\begin{aligned} P(SFFFF) &= P(S)P(F)P(F)P(F)P(F) \\ &= (0.07)(0.93)(0.93)(0.93)(0.93) \\ &= (0.07)(0.93)^4 \\ &= 0.05236 \end{aligned}$$

The probability calculation will be the same for any outcome with only one success ($x = 1$). It does not matter where in the sequence the single success occurs. It follows that

$$\begin{aligned} p(1) &= P(x = 1) \\ &= P(SFFFF \text{ or } FSFFF \text{ or } FFSFF \text{ or } FFFSF \text{ or } FFFFS) \\ &= 0.05236 + 0.05236 + 0.05236 + 0.05236 + 0.05236 \\ &= (5)(0.05236) \\ &= 0.26182 \end{aligned}$$

Similarly, there are 10 outcomes for which $x = 2$, because there are 10 ways to select two from among the five trials to be the S's: $SSFFF$, $SFSFF$, . . . , and $FFFSS$. The probability of each results from multiplying together (0.07) two times and (0.93) three times. For example,

$$\begin{aligned} P(SSFFF) &= (.07)(0.07)(0.93)(0.93)(0.93) \\ &= (0.07)^2(0.93)^3 \\ &= 0.00394 \end{aligned}$$

and so

$$\begin{aligned} p(2) &= P(x = 2) \\ &= P(SSFFF) + \cdots + P(FFFSS) \\ &= (10)(0.07)^2(0.93)^3 \\ &= 0.03941 \end{aligned}$$

The general form of the formula for calculating the probabilities associated with the different possible values of x is

$$\begin{aligned} p(x) &= P(x\ S\text{'}s \text{ among the five trials}) \\ &= (\text{number of outcomes with } x\ S\text{'}s)\cdot(\text{probability of any given outcome with } x\ S\text{'}s) \\ &= (\textit{number of outcomes with } x\ S\text{'}s)\cdot(0.07)^x(0.93)^{5-x} \end{aligned}$$

This form was seen previously where $p(2) = 10(0.07)^2(0.93)^3$.

The letter n is used to denote the number of trials in the binomial experiment. Then the number of outcomes with x S's is the number of ways of selecting x from among the n trials to be the success trials. A simple expression for this quantity is

$$\text{number of outcomes with } x \text{ successes} = \frac{n!}{x!(n-x)!}$$

where, for any positive whole number m, the symbol $m!$ (read "m factorial") is defined by

$$m! = m(m-1)(m-2)\cdots(2)(1)$$

and $0! = 1$.

The Binomial Distribution

Notation:

n = number of independent trials in a binomial experiment

p = probability of success for each trial

Then

$$\begin{aligned} p(x) &= P(x \text{ successes among } n \text{ trials}) \\ &= \frac{n!}{x!(n-x)!}p^x(1-p)^{n-x} \qquad x = 0, 1, 2, \ldots, n \end{aligned}$$

(continued)

Copyright 2025 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s). Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

The expressions $\binom{n}{x}$ or ${}_nC_x$ are sometimes used in place of $\frac{n!}{x!(n-x)!}$. Both are read as "n choose x" and represent the number of ways of choosing x items from a set of n. Using this notation, the binomial probability function can also be written as

$$p(x) = \binom{n}{x} p^x(1-p)^{n-x} \qquad x = 0, 1, 2, \ldots, n$$

or

$$p(x) = {}_nC_x\, p^x(1-p)^{n-x} \qquad x = 0, 1, 2, \ldots, n$$

Notice that here the binomial probability distribution is specified using a formula that allows calculation of the various probabilities rather than by giving a table or a probability histogram.

Example 7.18 Loss of Smell with COVID-19

Understand the context ➤

Loss of the ability to smell is one of the more common symptoms of COVID-19. The paper **"Prevalence and 6-Month Recovery of Olfactory Dysfunction: A Multicentre Study of 1363 COVID-19 Patients" (*Journal of Internal Medicine* [2021]: 451–461)** reported that about 86% of people who have COVID-19 lose some or all of their ability to smell and that about 5% of the patients who lose the ability to smell do not fully recover their sense of smell even after 6 months. Consider the random variable

x = number of patients who lose their sense of smell in a group of 20 COVID-19 patients

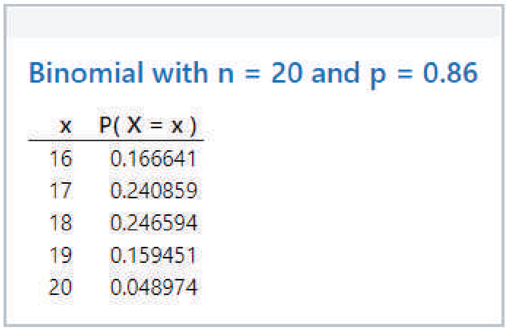
Observing whether or not a patient lost their sense of smell for each of these 20 patients can be considered as 20 trials, and the variable x counts the number of successes (loss of smell) in 20 trials. Assuming that the percentages reported in the paper are correct, x has a binomial distribution with $n = 20$ and $p = 0.86$.

Formulate a plan ➤

While you could use the binomial formula to calculate various probabilities for this random variable, using the formula can be tedious unless n is small. For this reason, it is much more common to rely on technology, such as a statistics software package, apps, or a graphing calculator to calculate binomial probabilities. For example, suppose you wanted to calculate the probability that more than 15 of the 20 patients experienced loss of smell. To use the binomial formula, you would have to calculate $p(16)$, $p(17)$, $p(18)$, $p(19)$, and $p(20)$ and then add these probabilities together. While you could do this, that is a lot of arithmetic!

Do the work ➤

Minitab and other software packages and graphing calculators will calculate binomial probabilities. Minitab was used to calculate the probabilities shown in the output below.



```
Binomial with n = 20 and p = 0.86

 x   P( X = x )
16     0.166641
17     0.240859
18     0.246594
19     0.159451
20     0.048974
```

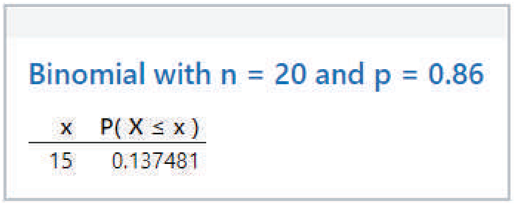
To calculate $P(x > 15)$, you can use these probabilities to obtain

$$P(x > 15) = 0.166641 + 0.240859 + 0.246594 + 0.159451 + 0.048974 = 0.86252$$

Copyright 2025 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s). Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

This is the probability of observing more than 15 people who experienced loss of smell in a group of 20 COVID-19 patients.

Minitab will also calculate cumulative probabilities, which are probabilities of the form $P(x \leq k)$, where k is some number between 0 and n. Another way to calculate $P(x > 15)$ is to determine $P(x \leq 15)$ and then subtract this number from 1, since $x \leq 15$ is the complement of $x > 15$. Minitab computed $P(x \leq 15)$, as shown in the output below.



```
Binomial with n = 20 and p = 0.86

 x   P( X ≤ x )
15    0.137481
```

Then

$$P(x > 15) = 1 - P(x \leq 15) = 1 - 0.137481 = 0.86252$$

You can also calculate other probabilities. For example:

$$P(x = 15) = 0.08681$$
$$P(14 \leq x \leq 17) = 0.52964$$

Interpret the results ➤

Based on these probabilities, you would say that in the long run, about 86.3% of all groups of 20 COVID-19 patients would include more than 15 who experienced loss of smell, about 8.7% of all groups of 20 would have exactly 15 who experienced loss of smell, and that just over half of all groups of 20 would have between 14 and 17 (inclusive) who experienced loss of smell.

Example 7.19 Computer Sales

Understand the context ➤

Suppose that 60% of all smartphones sold at an electronics store are iPhones and 40% are other models. The type of smartphone (iPhone or other) purchased by each of the next 12 smartphone customers will be noted. Define a random variable x as

x = number of smartphones among these 12 that are iPhones

What Do You Think?

With the random variable x defined this way, would you want to consider a customer who purchases an iPhone as a success or a failure? Why?

Formulate a plan ➤

Because x counts the number of iPhones, you can use S to denote the sale of a iPhone. Then x is a binomial random variable with $n = 12$ and $p = P(\text{S}) = 0.60$. The probability distribution of x is given by

$$p(x) = \frac{12!}{x!(12-x)!}(0.6)^x(0.4)^{12-x} \qquad x = 0, 1, 2, \ldots, 12$$

Do the work ➤

The probability that exactly four of the phones sold are iPhones is

$$\begin{aligned} p(4) &= P(x = 4) \\ &= \frac{12!}{4!8!}(0.6)^4(0.4)^8 \\ &= (495)(0.6)^4(0.4)^8 \\ &= 0.042 \end{aligned}$$

If group after group of 12 smartphone purchases is examined, the long-run percentage of those with exactly four iPhones will be 4.2%.

Copyright 2025 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s). Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

The probability that between four and seven (inclusive) are iPhones is

$$P(4 \leq x \leq 7) = P(x = 4 \text{ or } x = 5 \text{ or } x = 6 \text{ or } x = 7)$$

Since these outcomes are mutually exclusive, this is equal to

$$\begin{aligned} P(4 \leq x \leq 7) &= p(4) + p(5) + p(6) + p(7) \\ &= \frac{12!}{4!8!}(0.6)^4(0.4)^8 + \cdots + \frac{12!}{7!5!}(0.6)^7(0.4)^5 \\ &= 0.042 + 0.101 + 0.177 + 0.227 \\ &= 0.547 \end{aligned}$$

Notice that

$$\begin{aligned} P(4 < x < 7) &= P(x = 5 \text{ or } x = 6) \\ &= p(5) + p(6) \\ &= 0.278 \end{aligned}$$

so the probability depends on whether $<$ or $\leq$ appears in the inequality. (This is typical of *discrete* random variables.)

The binomial distribution formula can be tedious to use unless n is small. Statistical software and most graphing calculators can compute binomial probabilities. If you don't have access to technology, Appendix Table 9 can be used to find binomial probabilities for selected values of n and p.

Using Appendix Table 9

To find $p(x)$ for any particular value of x,

1. Locate the part of the table corresponding to your value of n (5, 10, 15, 20, or 25).
2. Move down to the row labeled with your value of x.
3. Go across to the column headed by the specified value of p.

The desired probability is at the intersection of the designated x row and p column. For example, when $n = 20$ and $p = 0.8$,

$$p(15) = P(x = 15) = (\text{entry at intersection of } n = 15 \text{ row and } p = 0.8 \text{ column}) = 0.175$$

Although $p(x)$ is positive for every possible x value, many probabilities are zero when rounded to three decimal places, so they appear as 0.000 in the table.

Sampling Without Replacement

Usually, sampling is carried out without replacement. This means that once an element has been selected for the sample, it is not a candidate for future selection. However, if sampling is done by selecting an element from the population, observing whether it is a success or a failure, and then returning it to the population before the next selection is made, the variable

x = number of successes observed in the sample

would fit all the requirements of a binomial random variable.

When sampling is done without replacement, the trials (individual selections) are not independent. In this case, the number of successes observed in the sample does not have a binomial distribution but rather a different type of distribution called a *hypergeometric distribution*. The probability calculations for this distribution are even more tedious than for the binomial distribution. Fortunately, when the sample size n is much smaller than N, the population size, probabilities calculated using the binomial distribution and those calculated using the hypergeometric distribution are very close in value. They are so close, in fact, that the difference is often ignored and binomial probabilities are used in place of the hypergeometric probabilities. The following box provides a guideline for determining whether it is appropriate to use the binomial probability distribution when sampling without replacement.

Let x denote the number of S's in a sample of size n selected without replacement from a population consisting of N individuals or objects. If $(n/N) \leq 0.05$ (that is, at most 5% of the population is sampled), then the binomial distribution provides a reasonable approximation to the actual probability distribution of x.*

Example 7.20 Video Game Use by Tweens and Teens

Understand the context ➤ The report **"The Common Sense Census: Media Use By Tweens and Teens" (Common Sense Media, 2019)** used data from a large representative sample to estimate that 64% of American between the ages of 8 and 18 play video games daily. Suppose that this percentage is correct for this group of American tweens and teens. Consider a random sample of $n = 20$ tweens and teens (much less than 5% of the population). Then

x = number in the sample who play video games daily

has (approximately) a binomial distribution with $n = 20$ and $p = 0.64$.

Do the work ➤ The probability that 12 of those in the sample play video games daily is (using technology)

$$p(12) = P(x = 12) = 0.16782$$

The probability that at least 18 in the sample play video games daily is

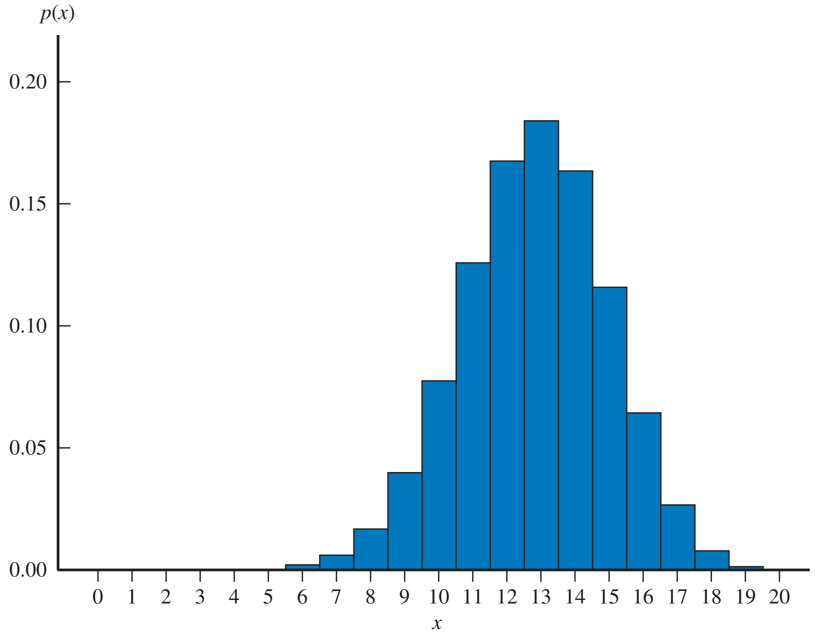
$$P(x \geq 18) = p(18) + p(19) + p(20) = 0.00962$$

Interpret the results ➤ If $p = 0.64$, only about 0.9% (just less than 1%) of all random samples of size 20 would include 18 or more who play video games daily. Because this probability is so small, if $x \geq 18$ were actually observed, you would have to wonder whether the reported value of $p = 0.64$ is correct.

Although it is possible that you would observe $x \geq 18$ when $p = 0.64$ (this would happen about 0.9% of the time in the long run), it might also be the case that p is actually greater than 0.64. In Chapter 10, you will see how hypothesis-testing methods can be used to decide which of two contradictory claims about a population (such as $p = 0.64$ or $p > 0.64$) is more believable.

Technology, the binomial formula or tables can be used to find each of the 21 probabilities $p(0)$, $p(1)$, ..., $p(20)$. Figure 7.15 shows the probability histogram for the binomial

Figure 7.15
The binomial probability histogram when $n = 20$ and $p = 0.64$.



---

*In Chapter 8, you will see a different situation where a similar condition is introduced, but where the requirement is that at most 10% of the population is included in the sample. Be careful not to confuse these two rules.

Copyright 2025 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s). Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

distribution with $n = 20$ and $p = 0.64$. Notice that the distribution is skewed to the left. (The binomial distribution is symmetric only when $p = 0.5$.)

Mean and Standard Deviation of a Binomial Random Variable

A binomial random variable x based on n trials has possible values 0, 1, 2, ..., n, so the mean value is

$$\mu_x = \Sigma xp(x) = (0)p(0) + (1)p(1) + \cdots + (n)p(n)$$

and the variance of x is

$$\begin{aligned}\sigma_x^2 &= \Sigma(x - \mu_x)^2 \cdot p(x) \\ &= (0 - \mu_x)^2 p(0) + (1 - \mu_x)^2 p(1) + \cdots + (n - \mu_x)^2 p(n)\end{aligned}$$

These expressions would be very tedious to evaluate for any particular values of n and p. Fortunately, there are simple formulas for the mean and standard deviation of a binomial random variable.

The mean value and the standard deviation of a binomial random variable x are

$$\mu_x = np \quad \text{and} \quad \sigma_x = \sqrt{np(1-p)}$$

Example 7.21 Budgets and Tracking Spending

Understand the context ➤

The report **"The 2019 Consumer Financial Literacy Survey" (The National Foundation for Credit Counseling, nfcc.org)** estimates that 40% of adult Americans say that they have a budget and keep track of their spending. This estimate was based on a representative sample of more than 2000 people. Suppose that 40% of all adult Americans have a budget and keep track of their spending. A random sample of $n = 25$ adult Americans is to be selected. Consider the random variable

x = number in the sample who have a budget and keep track of spending.

Formulate a plan ➤

Even though sampling is done without replacement, the sample size $n = 25$ is very small compared to the total number of adult Americans, so you can approximate the probability distribution of x using a binomial distribution with $n = 25$ and $p = 0.4$. Having a budget and keeping track of spending is identified as a success because this is the outcome counted by the random variable x.

Do the work ➤

The mean value of x is then

$$\mu_x = np = 25(0.40) = 10.0$$

and the standard deviation is

$$\sigma_x = \sqrt{np(1-p)} = \sqrt{25(0.40)(0.60)} = \sqrt{6} = 2.45$$

The probability that x is farther than 1 standard deviation from its mean value is

$$\begin{aligned}P(x < \mu_x - \sigma_x \textit{ or } x > \mu_x + \sigma_x) &= P(x < 7.55 \textit{ or } x > 12.45) \\ &= P(x \le 7) + P(x \ge 13) \\ &= p(0) + \cdots + p(7) + p(13) + \cdots + p(25) \\ &= 0.307 \quad \text{(using technology or Appendix Table 9)}\end{aligned}$$

The value of σ_x is 0 when $p = 0$ or $p = 1$. In these two cases, there is no uncertainty in x. You are sure to observe $x = 0$ when $p = 0$ and $x = n$ when $p = 1$. It is also easily verified that $p(1 - p)$ is largest when $p = 0.5$. This means that the binomial distribution spreads out the most when sampling from a 50–50 population. The farther p is from 0.5, the less spread out and the more skewed is the binomial distribution.

Copyright 2025 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s). Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

Geometric Distributions

A binomial random variable is defined as the number of successes in n independent trials, where each trial can result in either a success or a failure and the probability of success is the same for each trial. Suppose, however, that you are not interested in the number of successes in a fixed number of trials but rather in the number of trials that must be carried out before a success occurs. Two examples are counting the number of boxes of cereal that must be purchased before finding one with a rare toy and counting the number of games that a professional bowler must play before achieving a score over 250.

The variable

x = number of trials to first success

is called a **geometric random variable**, and the probability distribution that describes its behavior is called a **geometric probability distribution**.

Suppose an experiment consists of a sequence of trials that meet the following conditions:

1. The trials are independent.
2. Each trial can result in one of two possible outcomes, success and failure.
3. The probability of success is the same for all trials.

A **geometric random variable** is defined as

x = number of trials until the first success is observed (including the success trial)

The probability distribution of x is called the **geometric probability distribution**.

What Do You Think?

What aspect of a geometric experiment makes it different from a binomial experiment?

For example, suppose that 40% of the students who drive to campus at your university carry jumper cables. Your car has a dead battery and you don't have jumper cables, so you decide to stop students who are headed to the parking lot and ask them whether they have a pair of jumper cables. You might be interested in the number of students you would have to stop before finding one who has jumper cables.

FabrikaSimf/Shutterstock.com

If you define success as a student with jumper cables, a trial would consist of asking an individual student for help. The random variable

x = number of students who must be stopped before finding one with jumper cables

is an example of a geometric random variable, because it can be viewed as the number of trials to the first success in a sequence of independent trials.

The probability distribution of a geometric random variable is easy to construct. The letter p is used to denote the probability of success on any given trial. Possible outcomes can be denoted as follows:

Outcome	x = Number of Trials to First Success
S	1
FS	2
FFS	3
⋮	⋮
$FFFFFFS$	7
⋮	⋮

Copyright 2025 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s). Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

Each possible outcome consists of 0 or more failures followed by a single success. So,

$$\begin{aligned} p(x) &= P(x \text{ trials to first success}) \\ &= P(\underset{\substack{\uparrow \\ x-1 \text{ failures followed by a success on trial } x}}{FF\ldots FS}) \end{aligned}$$

Because the probability of success is p for each trial, the probability of failure for each trial is $1 - p$. Because the trials are independent,

$$\begin{aligned} p(x) &= P(x \text{ trials to first success}) = P(FF\ldots FS) \\ &= P(F)P(F)\cdots P(F)P(S) \\ &= (1-p)(1-p)\cdots(1-p)p \\ &= (1-p)^{x-1}p \end{aligned}$$

This leads us to the formula for the geometric probability distribution.

Geometric Probability Distribution

If x is a geometric random variable with probability of success $= p$ for each trial, then

$$p(x) = (1-p)^{x-1}p \quad x = 1, 2, 3, \ldots$$

Example 7.22 Jumper Cables

Consider the jumper cable scenario described previously. For this problem, $p = 0.4$, because 40% of the students who drive to campus carry jumper cables. The probability distribution of

x = number of students who must be stopped before finding a student with jumper cables

is

$$p(x) = (0.6)^{x-1}(0.4) \quad x = 1, 2, 3, \ldots$$

The probability distribution can now be used to calculate various probabilities. For example, the probability that the first student stopped has jumper cables (that is, $x = 1$) is

$$p(1) = (0.6)^{1-1}(0.4) = (0.6)^0(0.4) = 0.4$$

The probability that three or fewer students must be stopped is

$$\begin{aligned} P(x \le 3) &= p(1) + p(2) + p(3) \\ &= (0.6)^0(0.4) + (0.6)^1(0.4) + (0.6)^2(0.4) \\ &= 0.4 + 0.24 + 0.144 \\ &= 0.784 \end{aligned}$$

Exercises 7.51–7.69

7.51 CBS News reported that 4% of adult Americans have a food allergy **(June 1, 2017, cbsnews.com)**. Consider selecting 10 adult Americans at random. Define the random variable x as

x = number of people in the sample of 10 that have a food allergy.

Find the following probabilities. (Hint: See Examples 7.19 and 7.20.)

a. $p(x < 3)$
b. $p(x \le 3)$
c. $p(x \ge 4)$
d. $p(1 \le x \le 3)$

7.52 Examination of a large number of medical records found that only 82% of patients whose pregnancy was covered by Medicaid received what is considered to be routine follow-up care after delivery of their baby **("The Effect of Continuous Versus Pregnancy-Only Medicaid Eligibility on Routine Postpartum Care in Wisconsin, 2011–2015," *Maternal and Child Health Journal* [2020]: 1138–1150).** Suppose that a random sample of four patients whose pregnancy was covered by Medicaid is selected. Consider the random variable x = number in the sample who received routine follow-up care.

a. What are the possible values of x?

Copyright 2025 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s). Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

b. Is it reasonable to use a binomial distribution to model the behavior of x?

c. For each possible value of x, find the associated probability $p(x)$ and display the possible x values and $p(x)$ values in a table. (Hint: See Example 7.18.)

d. Construct a histogram displaying the probability distribution of x.

7.53 The article **"Should You Report That Fender-Bender?" (*Consumer Reports*, 2013:15)** reported that 7 in 10 auto accidents involve a single vehicle. Suppose 15 accidents are randomly selected. (Hint: See Examples 7.19 and 7.20.)

a. What is the probability that exactly four involve a single vehicle?

b. What is the probability that at most four involve a single vehicle?

c. What is the probability that exactly six involve multiple vehicles?

7.54 In a survey of a large representative sample of Americans who had booked air travel for the summer of 2021, it was reported that approximately 70% booked their flight online directly through the airline website **(Deloitte, May 25, 2021)**. Suppose that the actual percentage of Americans who book airline travel online directly through the airline website is 70%. Consider randomly selecting six Americans who have booked air travel and define the random variable x = number of the six selected passengers who booked online directly with the airline.

a. Calculate $p(4)$, and interpret this probability.

b. Calculate $p(6)$, the probability that all six booked flights online directly with the airline.

c. Calculate $P(x \geq 4)$.

7.55 Refer to the previous exercise, and suppose that 10 rather than six passengers are selected. Find the following:

a. $p(8)$

b. $P(x \leq 7)$

c. The probability that more than half booked flights online directly with the airline

7.56 Suppose that 25% of the customers of a grocery store use an express checkout. Consider five randomly selected customers, and let x denote the number among the five who use the express checkout.

a. Calculate $p(2)$.

b. Calculate $P(x \leq 1)$.

c. Calculate $P(2 \leq x)$. (Hint: Make use of your answer from Part (b).)

d. Calculate $P(x \neq 2)$.

7.57 According to a **Google/Harris Poll (February 2019)** of a representative sample of 3000 adults living in the United States, 13% reported that they reuse the same password for all their online accounts, which exposes them to greater risk of identity theft. Suppose that the true proportion of adults living in the United States who use the same password for all online accounts is 0.13 and that 20 adults living in the United States are selected at random.

a. What is the probability that more than three use the same password for all online accounts?

b. Which is more likely—that two or fewer use the same password for all online accounts or that more than two use the same password for all online accounts?

7.58 A breeder of show dogs is interested in the number of female puppies in a litter. If a birth is equally likely to result in a male or a female puppy, give the probability distribution of the variable x = number of female puppies in a litter of size 5.

7.59 *Women's Health Magazine* has an ongoing online survey asking people if they wash their sheets at least once a week **("How Often Should You Wash Your Sheets? Experts Weigh In," October 20, 2020)**. Even though microbiologists recommend that you wash your sheets at least once a week, the survey results as of November 17, 2021, were that only 49% of the survey participants said that they wash their sheets that often. Suppose the survey participants are representative of adult Americans and define the random variable x to be the number of adult Americans you would have to ask before you found someone who washes their sheets at least once a week.

a. Is the probability distribution of x binomial or geometric? Explain.

b. What is the probability that you would have to ask three people before finding one who washes sheets at least once a week?

c. What is the probability that fewer than four people would have to be asked before finding one who washes sheets at least once a week?

d. What is the probability that more than three people would have to be asked before finding one who washes sheets at least once a week?

7.60 Industrial quality control programs often include inspection of incoming parts from suppliers. If parts are purchased in large lots, a typical plan might be to select 20 parts at random from a lot and inspect them. Suppose that a lot is considered to be acceptable if one or fewer defective parts are found among those inspected. Otherwise, the lot is rejected and returned to the supplier. Use technology or Appendix Table 9 to find the probability of accepting lots that have each of the following (Hint: Identify success with a defective part):

a. 5% defective parts

b. 10% defective parts

c. 20% defective parts

Copyright 2025 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s). Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

7.61 Suppose that the probability is 0.1 that any given citrus tree will show measurable damage when the temperature falls to 30°F. (Hint: See Example 7.21.)

a. If the temperature does drop to 30°F, what is the expected number of citrus trees showing damage in orchards of 2000 trees?

b. What is the standard deviation of the number of trees that show damage?

7.62 Suppose that 30% of all automobiles undergoing an emissions inspection at an inspection station fail the inspection.

a. Among 15 randomly selected cars, what is the probability that at most 5 fail the inspection?

b. Among 15 randomly selected cars, what is the probability that between 5 and 10 (inclusive) fail the inspection?

c. Among 25 randomly selected cars, what is the mean value of the number that pass inspection, and what is the standard deviation of the number that pass inspection?

d. What is the probability that among 25 randomly selected cars, the number that pass is within 1 standard deviation of the mean value? (Hint: See Example 7.21.)

7.63 Suppose that you will take a multiple-choice exam consisting of 100 questions with five possible responses to each question. You have not studied and so must guess (select one of the five answers in a completely random fashion) on each question. Let x represent the number of correct responses on the test.

a. What kind of probability distribution does x have?

b. What is your expected score on the exam? (Hint: Your expected score is the mean value of the x distribution.)

c. Calculate the variance and standard deviation of x.

d. Based on your answers to Parts (b) and (c), is it likely that you would score over 50 on this exam? Explain the reasoning behind your answer.

7.64 Suppose that 20% of the 10,000 signatures on a recall petition are invalid. Would the number of invalid signatures in a sample of 2000 of these signatures have (approximately) a binomial distribution? Explain.

7.65 A city requires that smoke detectors be installed in all houses. There is concern that too many houses are still without detectors, so a costly inspection program is being considered. Let p be the proportion of all houses that have a smoke detector. A random sample of 25 houses is selected. If the sample strongly suggests that $p < 0.80$ (less than 80% have smoke detectors), as opposed to $p \geq 0.80$, the program will be implemented. Let x be the number of residences among the 25 that do have a detector, and consider the following decision rule: Reject the claim that $p \geq 0.8$ and implement the program if $x \leq 15$.

a. What is the probability that the program is implemented when $p = 0.80$?

b. What is the probability that the program is not implemented if $p = 0.70$?

c. What is the probability that the program is not implemented if $p = 0.60$?

d. How do the "error probabilities" of Parts (b) and (c) change if the value 15 in the decision rule is changed to 14?

7.66 Suppose that 90% of all registered California voters favor banning the release of information from exit polls in presidential elections until after the polls in California close. A random sample of 25 registered California voters is to be selected.

a. What is the probability that more than 20 favor the ban?

b. What is the probability that at least 20 favor the ban?

c. What are the mean value and standard deviation of the number of voters in the sample who favor the ban?

d. If fewer than 20 voters in the sample favor the ban, is this inconsistent with the claim that (at least) 90% of California registered voters favors the ban? (Hint: Consider $P(x < 20)$ when $p = 0.9$.)

7.67 Suppose a playlist on a music player consists of 100 songs, of which eight are by a particular artist. Songs are played by selecting a song at random (with replacement) from the playlist. Let the random variable x represent the number of songs played until a song by this artist is played.

a. Explain why the probability distribution of x is not binomial.

b. Find the following probabilities. (Hint: See Example 7.22.)

  i. $p(4)$
  ii. $P(x \leq 4)$
  iii. $P(x > 4)$
  iv. $P(x \geq 4)$

c. Interpret each of the probabilities in Part (b).

7.68 Sophie is a dog who loves to play catch. Unfortunately, she isn't very good, and the probability that she catches a ball is only 0.1. Let x be the number of tosses required until Sophie catches a ball.

a. Does x have a binomial or a geometric distribution?

b. What is the probability that it will take exactly two tosses for Sophie to catch a ball?

c. What is the probability that more than three tosses will be required?

Copyright 2025 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s). Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

7.69 Suppose that 5% of cereal boxes contain a prize and the other 95% contain the message, "Sorry, try again." Consider the random variable x, where x = number of boxes purchased until a prize is found.

a. What is the probability that at most two boxes must be purchased?

b. What is the probability that exactly four boxes must be purchased?

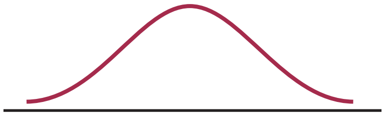
c. What is the probability that more than four boxes must be purchased?

Section 7.6 Normal Distributions

Normal distributions formalize the notion of mound-shaped histograms introduced in Chapter 4. Normal distributions are widely used for two reasons. First, they provide a reasonable approximation to the distribution of many different variables. They also play an important role in many of the inferential procedures that will be introduced in later chapters of this textbook.

Normal distributions are continuous probability distributions that are bell-shaped and symmetric, as shown in Figure 7.16. Normal distributions are sometimes referred to as *normal curves*.

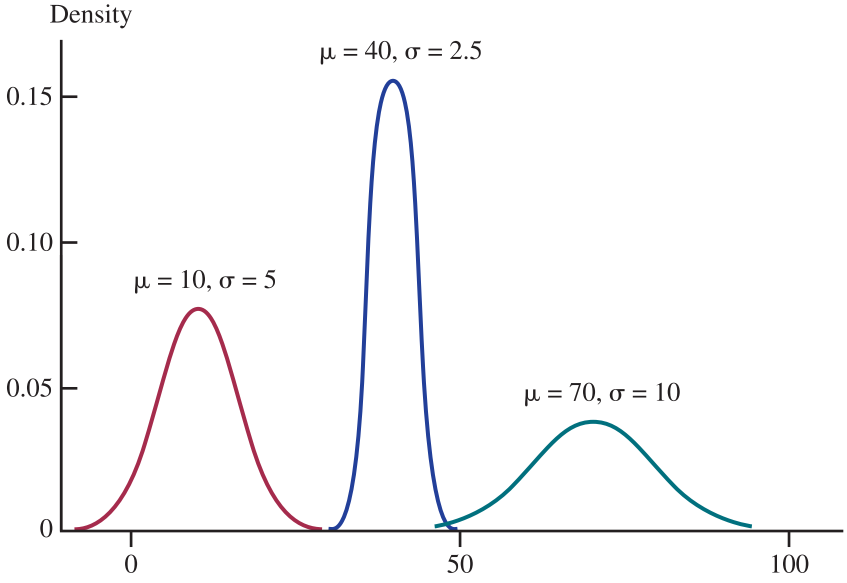
Figure 7.16
A normal distribution.



There are many different normal distributions, and they are distinguished from one another by their mean μ and standard deviation σ. The mean μ of a normal distribution describes where the corresponding curve is centered. The standard deviation σ describes how much the curve spreads out around that center. As with all continuous probability distributions, the total area under any normal curve is equal to 1.

Three normal distributions are shown in Figure 7.17. Notice that the smaller the standard deviation, the taller and narrower the corresponding curve. Remember that areas under a continuous probability distribution curve represent probabilities, so when the standard deviation is small, a larger area is concentrated near the center of the curve. This means that the chance of observing a value near the mean is much greater (because μ is at the center).

Figure 7.17
Three normal distributions.

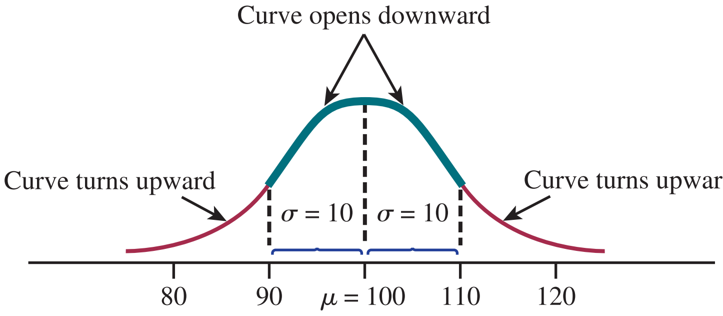


The value of μ is the number on the measurement axis lying directly below the top of the curve. The value of σ can be approximated from a picture of the curve. Consider the normal curve in Figure 7.18. Starting at the top (above $\mu = 100$) and moving to the right, the curve turns downward until it is above the value 110. After that point, it continues to decrease in height but is turning upward rather than downward. Similarly, to the left of

Copyright 2025 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s). Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

$\mu = 100$, the curve turns downward until it reaches 90 and then begins to turn upward. The curve changes from turning downward to turning upward at a distance of 10 on either side of μ. In general, σ is the distance to either side of μ at which a normal curve changes from turning downward to turning upward, so $\sigma = 10$ for the normal curve in Figure 7.18.

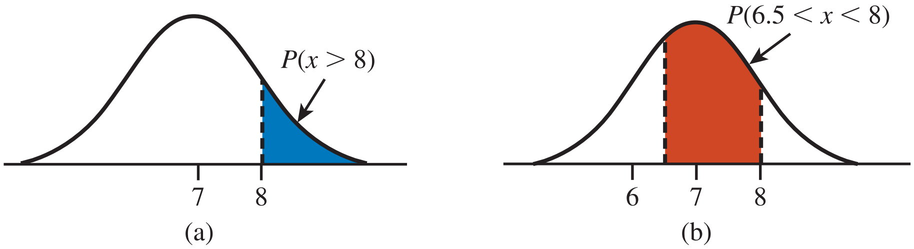
Figure 7.18
Mean μ and standard deviation σ for a normal curve.



When a normal distribution is used to describe the behavior of a random variable, a mean and a standard deviation must be specified. For example, a normal distribution with mean 7 and standard deviation 1 might be used as a model for the distribution of x = birth weight (in pounds). If this model is a reasonable description of the probability distribution, you can use areas under the normal curve with $\mu = 7$ and $\sigma = 1$ to approximate the distribution of the variable x = birth weight of a baby.

If this model is a reasonable description of the probability distribution of x, the probability that a birth weight is over 8 pounds (expressed symbolically as $P(x > 8)$) corresponds to the shaded area in Figure 7.19(a). The shaded area in Figure 7.19(b) represents the probability of a birth weight between 6.5 and 8 pounds, $P(6.5 < x < 8)$.

Figure 7.19
Normal distribution for birth weight:
(a) shaded area = $P(x > 8)$;
(b) shaded area = $P(6.5 < x < 8)$.



Unfortunately, calculating these probabilities (areas under a normal curve) is not simple. Fortunately, you can rely on technology or a table of areas for a reference normal distribution, called the **standard normal distribution**.

Definitions

Standard normal distribution: The normal distribution with

$$\mu = 0 \quad \text{and} \quad \sigma = 1$$

The corresponding density curve is called the **standard normal curve**.

The letter z is used to represent a variable whose distribution is described by the standard normal curve. The term ***z* curve** is often used in place of standard normal curve.

There aren't many naturally occurring variables with distributions that are well described by the standard normal distribution. However, this distribution is important because it can be used in probability calculations for other normal distributions. When you are interested in finding a probability based on some other normal curve, you can either rely on technology or translate the problem into an equivalent problem that involves finding an area under the standard normal curve. A table for the standard normal distribution can then be used to find the desired area. To be able to do this, you must first learn to work with the standard normal distribution.

For now, let's assume that you will be using technology (such as a statistical software package like Minitab or SALT, the Statistical Analysis and Learning Tool that accompanies this text) when you are working with normal distributions. You could also use a graphing calculator. For those who don't have access to technology, using a table for the standard normal distribution is covered in the optional subsection at the end of this section.

In working with normal distributions, you need two general skills:

1. You will need to use the normal distribution to calculate probabilities, which are areas under a normal curve and above given intervals.
2. You will need to be able to characterize extreme values in the distribution, such as the largest 5%, the smallest 1%, and the most extreme 5% (which would include the largest 2.5% and the smallest 2.5%).

All statistics software packages and most graphing calculators have normal distribution functions. In addition, many normal distribution apps are widely available. Examples 7.23–7.25 illustrate using Minitab, but other packages and apps have similar functionality.

Example 7.23 Normal Body Temperature

Understand the context ➤

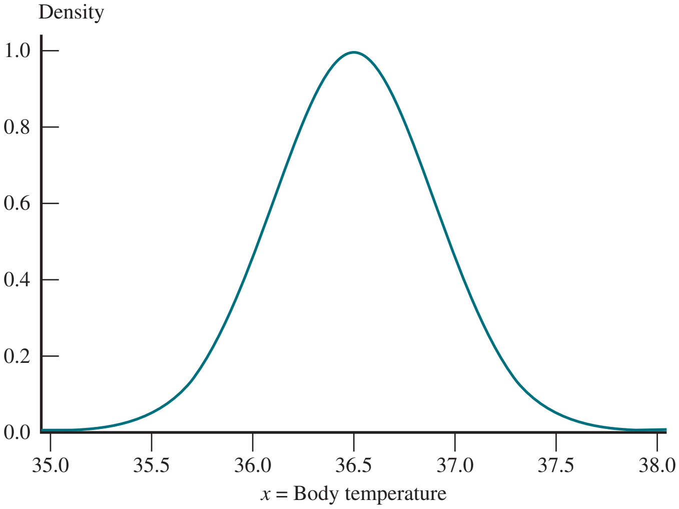
The long-standing definition (dating back to 1868) of normal temperature for healthy adults has been 37°C (98.6°F), but recent research has suggested that this is no longer accurate. The authors of the paper **"Using Smartphone Crowdsourcing to Redefine Normal and Febrile Temperatures in Adults: Results from the Feverprints Study" (*Journal of General Internal Medicine [2018]: 2046–2047*)** collected oral temperature data from a large number of adults who entered their temperatures into a smartphone app. Based on the data collected, a reasonable model for the probability distribution of

x = body temperature of a randomly selected healthy adult

is a normal distribution with a mean of 36.5°C (97.7°F) and a standard deviation of 0.4°C.

Figure 7.20 shows this normal distribution.

Figure 7.20
Normal distribution for body temperature (mean 36.5 and standard deviation 0.4)



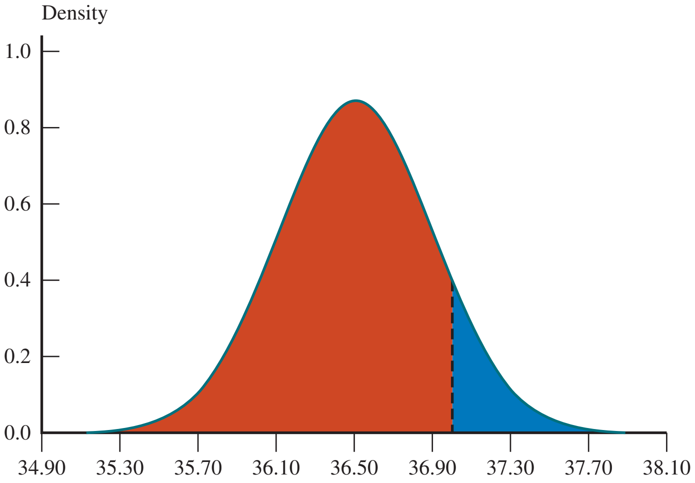
Consider the following questions:

1. What proportion of healthy adults have a body temperature that is less than the old standard of 37°C?

Copyright 2025 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s). Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

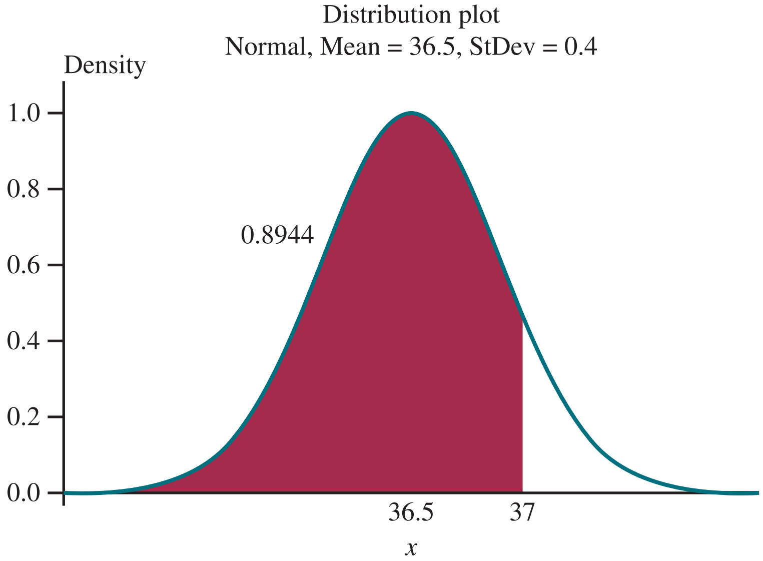
2. What proportion of healthy adults have a body temperature that is within 0.5°C of 36.5°C (between 36°C and 37°C)?
3. Some define a fever as a temperature that is above the 99th percentile of the normal temperature distribution. What temperatures would be considered fever?

Do the work ➤ Let's start with the first question. This is asking you to find $P(x < 37)$. Because x is a continuous random variable, this corresponds to an area under the normal curve and to the left of 37, as shown here:



You can find probability calculators in Minitab under Probability Distribution Plot on the Graph menu. To calculate this probability, select View Probability from the pop-up menu, and then select Normal from the drop-down menu of distributions and then specify the mean (36.5) and the standard deviation (0.4). You can then use the probability calculator to find the desired area by clicking on the Shaded Area tab, clicking on x value and left tail, and then entering 37 in the appropriate box. Figure 7.21 shows the output that results.

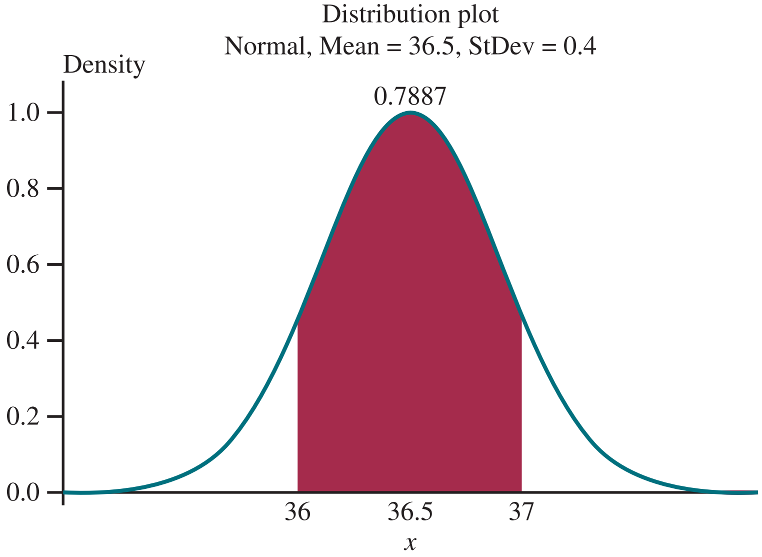
Figure 7.21
Output from Minitab $P(x < 37)$



From the output, $P(x < 37) = 0.8944$. According to this model, approximately 89.4% of healthy adults have a body temperature that is less than 37°C.

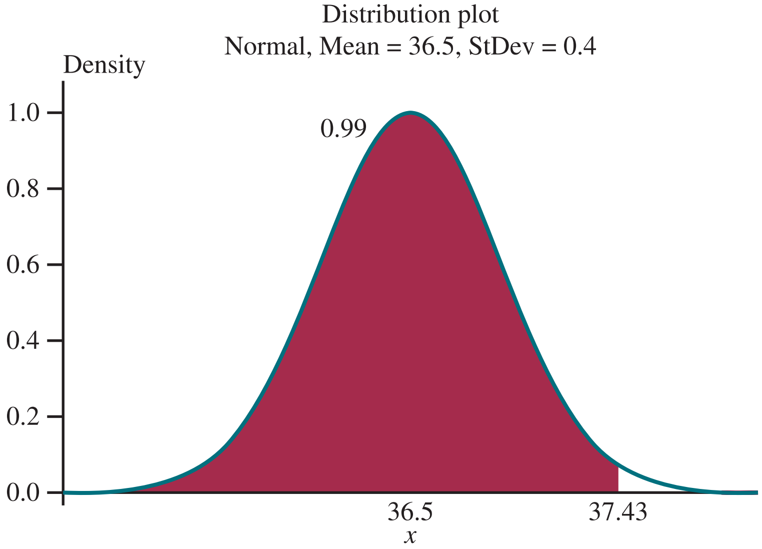
The second question posed was "What proportion of healthy adults have a body temperature that is within 0.5°C of 36.5 (between 36°C and 37°C)?" This is asking for $P(36 < x < 37)$. This corresponds to the area under the normal curve and above the interval from 36 to 37. Figure 7.22 shows Minitab output. Notice that "Middle" was selected because you are interested in the probability that x is between 36 and 37. From the output, $P(36 < x < 37) = 0.7887$. According to this model, approximately 78.9% of healthy adults have body temperatures between 36°C and 37°C.

Figure 7.22
Output from Minitab $P(36 < x < 37)$



Moving to the third question, you would need to decide what temperatures would be considered fever if fever is defined as a temperature that is above the 99th percentile of the normal temperature distribution. The temperature that is the 99th percentile is the temperature for which 99% of healthy adults would have a lower body temperature. In this situation, you want to find the temperature that would cut off an area of 0.99 under the curve and to the left (corresponding to 99% of the area below). Output from choosing the left tail area and specifying 0.99 as a left tail probability (rather than specifying an x value) is shown in Figure 7.23. From this output, you can see that the 99th percentile of the body temperature distribution is 37.43°C (which corresponds to 99.4°F). Only about 1% of healthy adults would have a body temperature that is greater than 37.43°C.

Figure 7.23
Output from Minitab 99th percentile



This example used Minitab to find an area to the left of a value, the area between two values, and a percentile for a given normal distribution. Minitab could also be used to calculate an area to the right of a value. Other statistics software packages and web apps have similar functionality and graphing calculators also calculate areas and percentiles for normal distributions.

Copyright 2025 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s). Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

Next, consider identifying extreme values. Suppose that you want to describe the values included in the largest 2% of a normal distribution or the values making up the most extreme 5% (which includes the largest 2.5% and the smallest 2.5%). Examples 7.24 and 7.25 illustrate how technology can be used to identify extreme values in normal distributions.

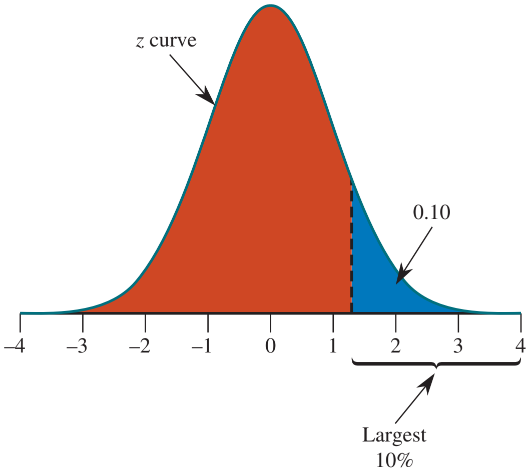
Example 7.24 Identifying Extreme Values

In later chapters, you will need to identify extreme values for the standard normal distribution. Recall that the standard normal distribution is the normal distribution with mean = 0 and standard deviation = 1. Suppose that you want to describe the values that make up the greatest 10% of a standard normal distribution. Because z is usually used to represent a variable that has a standard normal distribution, you can use z^* to represent the value that separates the largest 10% from the 90% that are smaller than this value. Symbolically, you are trying to find a value such that

$$P(z > z^*) = 0.10$$

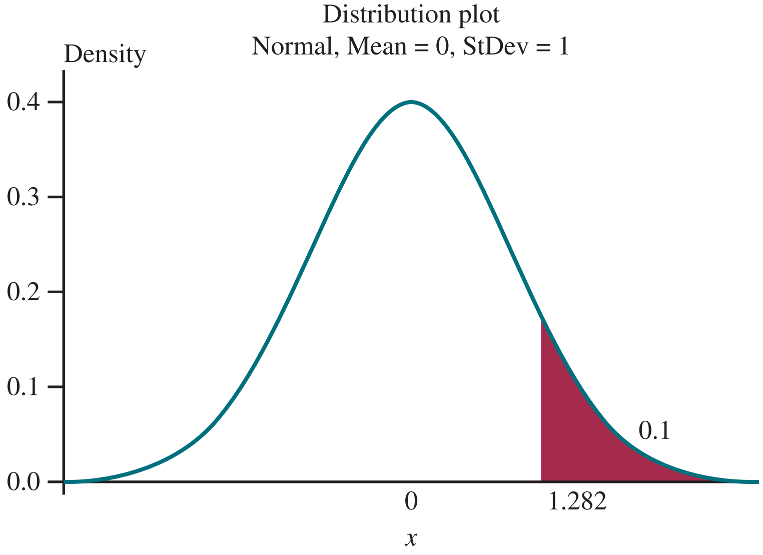
This is illustrated in Figure 7.24.

Figure 7.24
The largest 10% of the standard normal distribution



Using a software package like Minitab, you could find the value of z^* by either specifying an area to the right of 0.10 or an area to the left of 0.90. Output from Minitab is shown in Figure 7.25(a). From the output, $z = 1.282$. Values greater than 1.282 make up the largest 10% of the standard normal distribution.

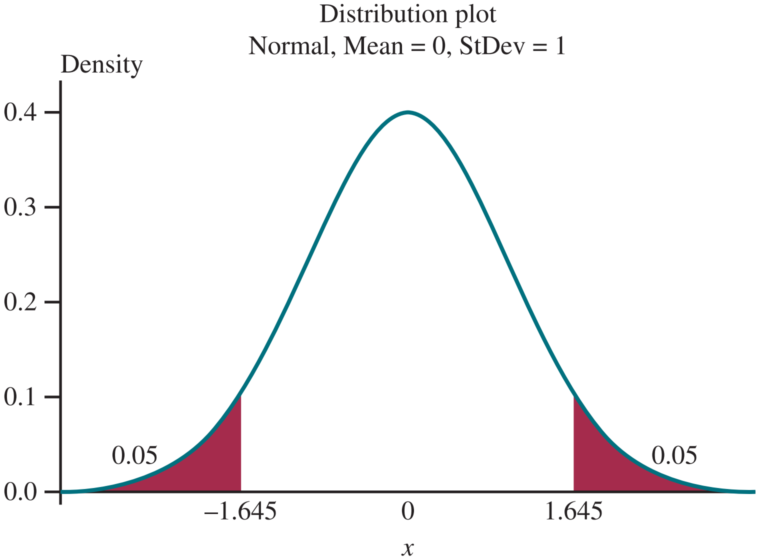
Figure 7.25
The largest 10% and the most extreme 10% of the standard normal distribution:
(a) Minitab Output for largest 10% of the standard normal distribution;



(a)

Copyright 2025 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s). Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

Figure 7.25
(b) Minitab Output for the most extreme 10% of the standard normal distribution



(b)

Sometimes you are interested in identifying the most extreme (largest *or* smallest) values in a distribution. For example, suppose you want to identify the values that describe the most extreme 10% of the standard normal distribution. Here you want to separate the middle 90% from the most extreme 10%. Because the normal distribution is symmetric, the most extreme 10% is equally divided between the high side and the low side of the distribution, resulting in an area of 0.05 for each of the tails of the z curve. Using Minitab (Figure 7.25(b)), you can see that values that are either less than –1.645 or greater than 1.645 are the most extreme 10% of the standard normal distribution.

Example 7.25 Tesla Batteries

Understand the context ➤

Tesla claims that the batteries that power its electric vehicles are designed to outlast the vehicle. In the **2020 Tesla Impact Report**, Tesla states that the battery averages 90% of its original capacity after having been driven 200,000 miles, which is longer than the typical lifetime of the car. Based on a graph given in the report, a reasonable model for the distribution of the random variable

x = percentage of original battery capacity after 200,000 miles

for Tesla vehicles is a normal distribution with a mean of 90% and a standard deviation of about 3%.

Consider the following questions:

1. Based on this model, what is the probability that a Tesla that has been driven 200,000 miles has a battery capacity that is less than 85% of its original battery capacity?
2. Would it be unusual for a Tesla that has been driven 200,000 miles to have more than 96% of its original battery capacity?
3. What battery capacities would describe the middle 50% of cars that have been driven 200,000 miles?

What Do You Think?

Based on the output given in Figure 7.26, how would you answer each of the questions just posed?

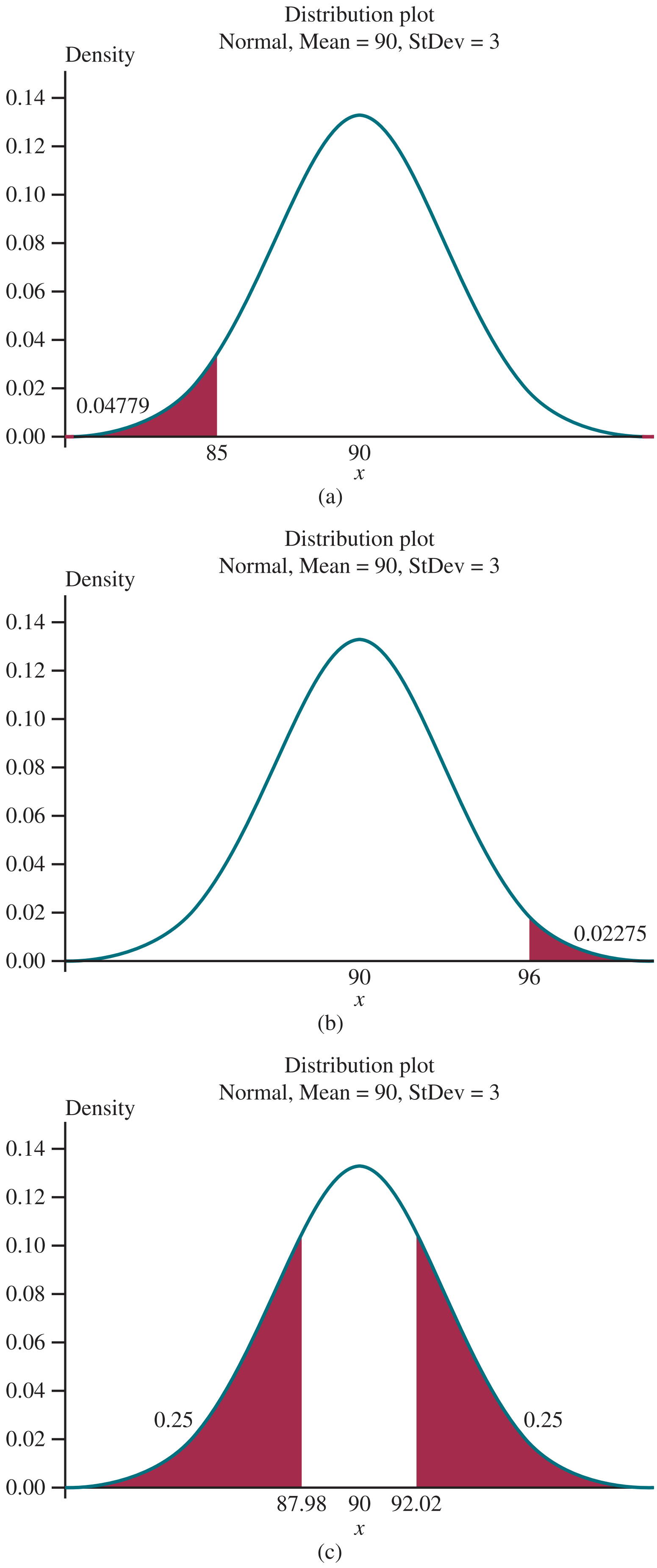
Do the work ➤

Minitab was used to answer these three questions, and the output is given in Figure 7.26(a)–(c).

1. The probability that a Tesla that has been driven 200,000 miles has a battery capacity that is less than 85% of its original battery capacity is $P(x < 85) = 0.04779$.

Copyright 2025 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s). Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

Figure 7.26
Minitab output for Example 7.25



2. The probability that a Tesla that has been driven 200,000 miles has a battery capacity that is greater than 96% of its original battery capacity is $P(x > 96) = 0.02275$, so it would be unusual to see a Tesla that had been driven 200,000 miles have 96% or more of its original battery capacity. Only about 2% of Teslas with 200,000 miles would have a battery capacity this large.
3. The Teslas that have been driven 200,000 miles that make up the middle 50% in terms of battery capacity have battery capacities that are between about 88% and 92% of the original battery capacity.

Other statistics software packages and web apps have similar functionality and graphing calculators could also be used to answer questions like the ones considered in this example.

Working with Standard Normal Tables (Optional)

If you are using technology like a statistics software package, web apps, or a graphing calculator, you won't need to use tables of normal curve areas. But if you don't have access to technology, tables provide an alternative. In this optional subsection, you will see how to use tables to do the following:

1. Use the normal distribution to calculate probabilities, which are areas under a normal curve and above given intervals.
2. Characterize extreme values in the distribution, such as the largest 5%, the smallest 1%, and the most extreme 5% (which would include the largest 2.5% and the smallest 2.5%).

Let's begin by looking at how to accomplish these tasks for the standard normal distribution.

The standard normal or z curve is shown in Figure 7.27(a). It is centered at $\mu = 0$, and the standard deviation, $\sigma = 1$, is a measure of how much it spreads out about its mean. Notice that this picture is consistent with the Empirical Rule of Chapter 4: About 95% of the area is associated with values that are within 2 standard deviations of the mean (between -2 and 2), and almost all of the area is associated with values that are within 3 standard deviations of the mean (between -3 and 3).

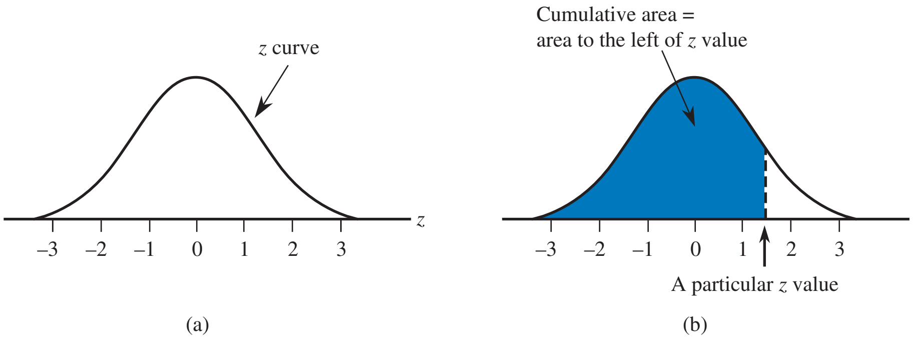


Figure 7.27
(a) The standard normal curve;
(b) a cumulative area

Appendix Table 2 gives cumulative z curve areas like the one shown in Figure 7.27(b) for many different values of z. The smallest value for which the cumulative area is given is $z = -3.89$, a value far out in the lower tail of the z curve. The next smallest value for which an area is given is -3.88, then -3.87, and so on in increments of 0.01 all the way up to 3.89.

Using the Table of Standard Normal Curve Areas

For any number z^* between -3.89 and 3.89 and rounded to two decimal places, Appendix Table 2 gives

$$(\text{area under } z \text{ curve to the left of } z^*) = P(z < z^*) = P(z \leq z^*)$$

(continued)

Copyright 2025 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s). Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

where the letter z is used to represent a random variable whose distribution is the standard normal distribution.

To find this probability using the table, locate the following:

1. The row labeled with the sign of z^* and the digit on either side of the decimal point (for example, -1.7 or 0.5)
2. The column identified with the second digit to the right of the decimal point in z^* (for example, .06 if $z^* = -1.76$)

The number at the intersection of this row and column is the desired probability, $P(z < z^*)$.

A portion of the table of standard normal curve areas appears in Figure 7.28. To find the area under the z curve to the left of 1.42, look in the row labeled 1.4 and the column labeled .02 (the highlighted row and column in Figure 7.28). From the table, the corresponding cumulative area is 0.9222. So,

$$z \text{ curve area to the left of } 1.42 = 0.9222$$

Figure 7.28
Portion of the table of standard normal curve areas

z^*	.00	.01	.02	.03	.04	.05
0.0	.5000	.5040	.5080	.5120	.5160	.5199
0.1	.5398	.5438	.5478	.5517	.5557	.5596
0.2	.5793	.5832	.5871	.5910	.5948	.5987
0.3	.6179	.6217	.6255	.6293	.6331	.6368
0.4	.6554	.6591	.6628	.6664	.6700	.6736
0.5	.6915	.6950	.6985	.7019	.7054	.7088
0.6	.7257	.7291	.7324	.7357	.7389	.7422
0.7	.7580	.7611	.7642	.7673	.7704	.7734
0.8	.7881	.7910	.7939	.7967	.7995	.8023
0.9	.8159	.8186	.8212	.8238	.8264	.8289
1.0	.8413	.8438	.8461	.8485	.8508	.8531
1.1	.8643	.8665	.8686	.8708	.8729	.8749
1.2	.8849	.8869	.8888	.8907	.8925	.8944
1.3	.9032	.9049	.9066	.9082	.9099	.9115
1.4	.9192	.9207	.9222	.9236	.9251	.9265
1.5	.9332	.9345	.9357	.9370	.9382	.9394
1.6	.9452	.9463	.9474	.9484	.9495	.9505
1.7	.9554	.9564	.9573	.9582	.9591	.9599
1.8	.9641	.9649	.9656	.9664	.9671	.9678

$P(z < 1.42)$

You can also use the table to find the area to the right of 1.42. Because the total area under the z curve is 1, it follows that

$$\begin{aligned}(z \text{ curve area to the right of } 1.42) &= 1 - (z \text{ curve area to the left of } 1.42)\\ &= 1 - 0.9222\\ &= 0.0778\end{aligned}$$

These probabilities can be interpreted to mean that about 92.22% of observed z values will be smaller than 1.42, and about 7.78% will be larger than 1.42.

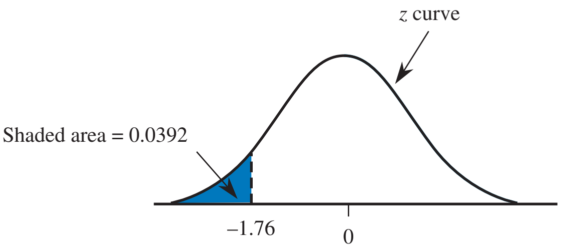
Example 7.26 Finding Standard Normal Curve Areas

The probability $P(z < -1.76)$ appears in the z table at the intersection of the -1.7 row and the .06 column. The result is

$$P(z < -1.76) = 0.0392$$

Copyright 2025 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s). Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

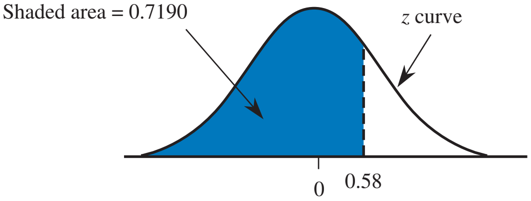
as shown in the following figure:



In other words, about 3.9% of observed z values will be smaller than -1.76. Similarly,

$$P(z \le 0.58) = \text{entry in 0.5 row and .08 column of Table 2} = 0.7190$$

as shown in the following figure:



Now consider $P(z < -4.12)$. This probability does not appear in Appendix Table 2; there is no -4.1 row. However, it must be less than $P(z < -3.89)$, the smallest z value in the table, because -4.12 is farther out in the lower tail of the z curve. Since $P(z < -3.89) \approx 0.0000$ (that is, zero to four decimal places), it follows that

$$P(z < -4.12) \approx 0$$

Similarly,

$$P(z < 4.12) > P(z < 3.89) \approx 1.0000$$

so

$$P(z < 4.12) \approx 1$$

As illustrated in Example 7.26, you can use Appendix Table 2 to calculate other probabilities involving z. The probability that z is larger than a value c is

$$P(z > c) = \text{area under the } z \text{ curve to the right of } c = 1 - P(z \le c)$$

In other words, the area to the right of a value (a right-tail area) is 1 minus the corresponding cumulative area. This is illustrated in Figure 7.29.

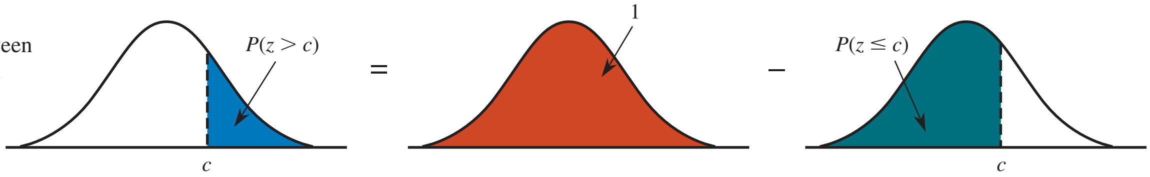


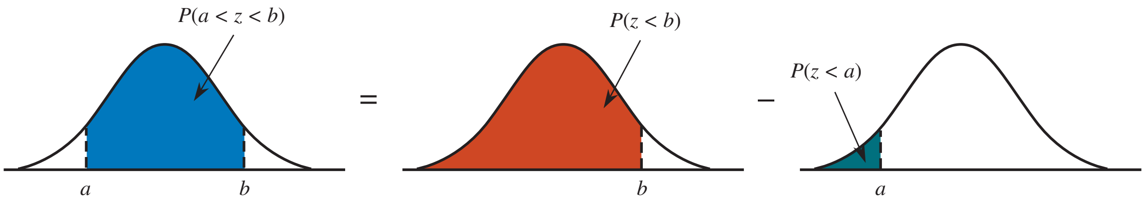
Figure 7.29 The relationship between an upper-tail area and a cumulative area

Similarly, the probability that z falls in the interval between a lower endpoint a and an upper endpoint b is

$$P(a < z < b) = \text{area under the } z \text{ curve and above the interval from } a \text{ to } b$$
$$= P(z < b) - P(z < a)$$

That is, $P(a < z < b)$ is the difference between two cumulative areas, as illustrated in Figure 7.30.

Figure 7.30
$P(a < z < b)$ as the difference between two cumulative areas

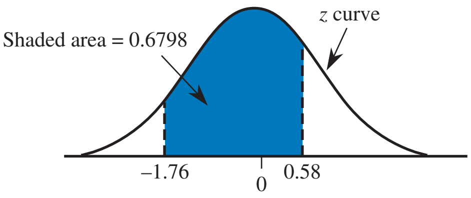


Example 7.27 More About Standard Normal Curve Areas

The probability that z is between -1.76 and 0.58 is

$$\begin{aligned} P(-1.76 < z < 0.58) &= P(z < 0.58) - P(z < -1.76) \\ &= 0.7190 - 0.0392 \\ &= 0.6798 \end{aligned}$$

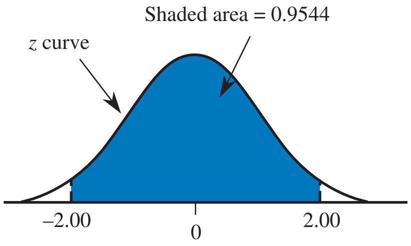
as shown in the following figure:



The probability that z is between -2 and $+2$ (within 2 standard deviations of its mean, since $\mu = 0$ and $\sigma = 1$) is

$$\begin{aligned} P(-2.00 < z < 2.00) &= P(z < 2.00) - P(z < -2.00) \\ &= 0.9772 - 0.0228 \\ &= 0.9544 \end{aligned}$$

as shown in the following figure:

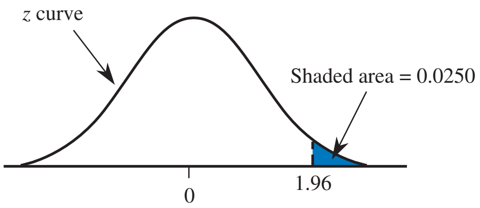


This last probability is the basis for one part of the Empirical Rule, which states that when a data distribution is well approximated by a normal curve, approximately 95% of the values are within 2 standard deviations of the mean.

The probability that the value of z exceeds 1.96 is

$$\begin{aligned} P(z > 1.96) &= 1 - P(z < 1.96) \\ &= 1 - 0.9750 \\ &= 0.0250 \end{aligned}$$

as shown in the following figure:



That is, 2.5% of the area under the z curve lies to the right of 1.96 in the upper tail.

Similarly,

$$\begin{aligned} P(z > -1.28) &= \text{area to the right of } -1.28 \\ &= 1 - P(z < -1.28) \\ &= 1 - 0.1003 \\ &= 0.8997 \\ &\approx 0.90 \end{aligned}$$

Identifying Extreme Values

Suppose that you want to describe the values included in the smallest 2% of a distribution or the values making up the most extreme 5% (which includes the largest 2.5% and the smallest 2.5%). Examples 7.28 and 7.29 show how to identify extreme values in the standard normal distribution.

Example 7.28 Identifying Extreme Values

Suppose that you want to describe the values that make up the smallest 2% of the standard normal distribution. Symbolically, you are trying to find a value (call it z^*), such that

$$P(z < z^*) = 0.02$$

This is illustrated in Figure 7.31, which shows that the cumulative area for z^* is 0.02.

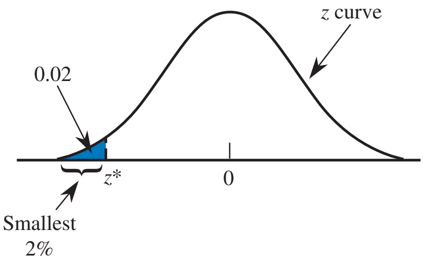


Figure 7.31
The smallest 2% of the standard normal distribution

This means you should look for a cumulative area of 0.0200 in the body of Appendix Table 2. The closest cumulative area in the table is 0.0202 (in the −2.0 row and .05 column), so $z^* = -2.05$. Variable values less than −2.05 make up the smallest 2% of the standard normal distribution. Notice that, because of the symmetry of the normal curve, variable values greater than 2.05 make up the largest 2% of the standard normal distribution.

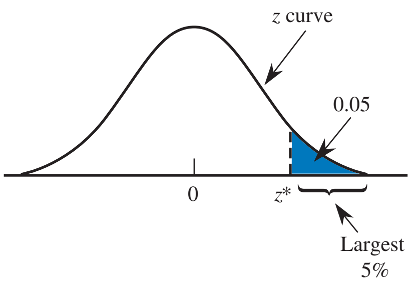
Now suppose that you are interested in the largest 5% of all z values. You would then be trying to find a value of z^* for which

$$P(z > z^*) = 0.05$$

as illustrated in Figure 7.32. Because Appendix Table 2 always works with cumulative area (area to the left), the first step is to determine

$$\text{area to the left of } z^* = 1 - 0.05 = 0.95$$

Figure 7.32
The largest 5% of the standard normal distribution



Looking for the cumulative area closest to 0.95 in Appendix Table 2, you find that 0.95 falls exactly halfway between 0.9495 (corresponding to a z value of 1.64) and 0.9505 (corresponding to a z value of 1.65). Because 0.9500 is exactly halfway between the two areas, use a z value that is halfway between 1.64 and 1.65. (If one value had been closer to 0.9500 than the other, you could just use the z value corresponding to the closest area.) This gives

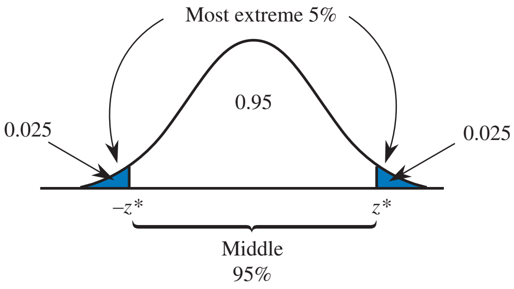
$$z^* = \frac{1.64 + 1.65}{2} = 1.645$$

Values greater than 1.645 make up the largest 5% of the standard normal distribution. By symmetry, -1.645 separates the smallest 5% of all z values from the others.

Example 7.29 More Extremes

Sometimes you are interested in identifying the most extreme (unusually large *or* small) values in a distribution. How can you describe the values that make up the most extreme 5% of the standard normal distribution? Here you want to separate the middle 95% from the most extreme 5%. This is illustrated in Figure 7.33.

Figure 7.33
The most extreme 5% of the standard normal distribution



Because the standard normal distribution is symmetric, the most extreme 5% is equally divided between the high side and the low side of the distribution, resulting in an area of 0.025 for each of the tails of the z curve. To find z^*, first determine the cumulative area for z^*, which is

$$\text{area to the left of } z^* = 0.95 + 0.025 = 0.975$$

The cumulative area 0.9750 appears in the 1.9 row and .06 column of Appendix Table 2, so $z^* = 1.96$. Symmetry about 0 implies that if z^* denotes the value that separates the largest 2.5%, the value that separates the smallest 2.5% is $-z^*$, or -1.96. For the standard normal distribution, values that are either less than -1.96 or greater than 1.96 are the most extreme 5%.

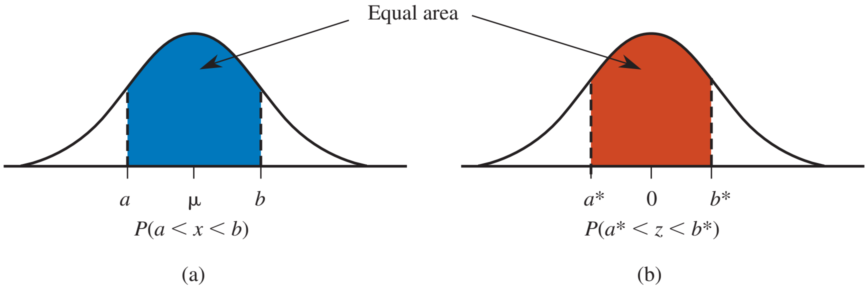
Other Normal Distributions

Areas under the z curve can be used to calculate probabilities and to describe values for *any* normal distribution. Remember that the letter z is usually reserved for variables that have a standard normal distribution. Other letters, such as x or y, are used to denote a variable whose distribution is described by a normal curve with mean μ and standard deviation σ.

Copyright 2025 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s). Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

Suppose that you want to calculate $P(a < x < b)$, the probability that the variable x lies in a particular range. This probability corresponds to an area under a normal curve and above the interval from a to b, as shown in Figure 7.34(a).

Figure 7.34
Equality of normal curve areas: (a) normal distribution with mean μ and standard deviation σ; (b) standard normal distribution



The strategy for obtaining this probability is to find an equivalent problem involving the standard normal distribution. Finding an equivalent problem means determining an interval (a^*, b^*) that has the same probability for z (same area under the z curve) as the interval (a, b) in the original normal distribution (see Figure 7.34(b)). The asterisk is used to distinguish a and b, the values from the original normal distribution from a^* and b^*, the values from the z distribution.

To find a^* and b^*, simply calculate z-scores for a and b. This process is called **standardizing** the endpoints. For example, suppose that the variable x has a normal distribution with mean $\mu = 100$ and standard deviation $\sigma = 5$. To find

$$P(98 < x < 107)$$

first translate this problem into an equivalent one for the standard normal distribution. Recall from Chapter 3 that a z-score is calculated by first subtracting the mean and then dividing by the standard deviation. Converting the lower endpoint $a = 98$ to a z-score gives

$$a^* = \frac{98 - 100}{5} = \frac{-2}{5} = -0.40$$

and converting the upper endpoint yields

$$b^* = \frac{107 - 100}{5} = \frac{7}{5} = 1.40$$

Then

$$P(98 < x < 107) = P(-0.40 < z < 1.40)$$

The probability $P(-0.40 < z < 1.40)$ can now be evaluated using technology or Appendix Table 2.

Finding Probabilities

To calculate probabilities for any normal distribution, standardize the relevant values and then use the table of z curve areas. More specifically, if x is a variable whose behavior is described by a normal distribution with mean μ and standard deviation σ, then

$$P(x < b) = P(z < b^*)$$
$$P(x > a) = P(z > a^*)$$
$$P(a < x < b) = P(a^* < z < b^*)$$

where z is a variable whose distribution is standard normal and

$$a^* = \frac{a - \mu}{\sigma} \qquad b^* = \frac{b - \mu}{\sigma}$$

Copyright 2025 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s). Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

Example 7.30 Newborn Birth Weights

Understand the context ➤

Data from the paper **"Birth Weight Curves Tailored to Maternal World Region" (*Journal of Obstetrics and Gynaecology Canada* [2012]: 159–171)** suggest that a normal distribution with mean $\mu = 3500$ grams and standard deviation $\sigma = 550$ grams is a reasonable model for the probability distribution of x = birth weight of a randomly selected full-term baby born in Canada. What proportion of birth weights in Canada are between 2900 and 4700 grams?

Do the work ➤

To answer this question, you must find

$$P(2900 < x < 4700)$$

First, translate the interval endpoints to equivalent endpoints for the standard normal distribution:

$$a^* = \frac{a - \mu}{\sigma} = \frac{2900 - 3500}{550} = -1.09$$

$$b^* = \frac{b - \mu}{\sigma} = \frac{4700 - 3500}{550} = 2.18$$

Then

$$\begin{aligned} P(2900 < x < 4700) &= P(-1.09 < z < 2.18) \\ &= (z \text{ curve area to the left of } 2.18) \\ &\quad - (z \text{ curve area to the left of } -1.09) \\ &= 0.9854 - 0.1379 \\ &= 0.8475 \end{aligned}$$

The probabilities for x and z are shown in Figure 7.35. If birth weights were observed for many babies born in Canada, about 85% of them would be between 2900 and 4700 grams.

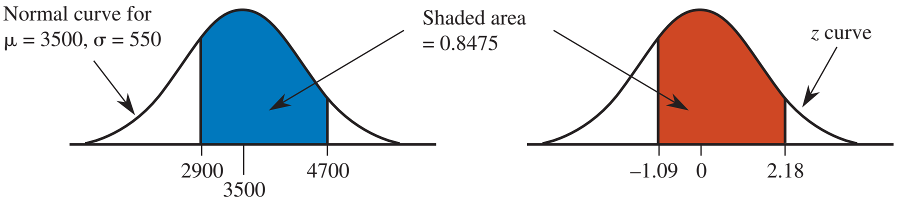


Figure 7.35 $P(2900 < x < 4700)$ and corresponding z curve area for the birth weight distribution of Example 7.30

What is the probability that a randomly chosen baby will have a birth weight greater than 4500? To evaluate $P(x > 4500)$. First calculate

$$a^* = \frac{a - \mu}{\sigma} = \frac{4500 - 3500}{550} = 1.82$$

Then (see Figure 7.36).

$$\begin{aligned} P(x > 4500) &= P(z > 1.82) \\ &= z \text{ curve area to the right of } 1.82 \\ &= 1 - (z \text{ curve area to the left of } 1.82) \\ &= 1 - 0.9656 \\ &= 0.0344 \end{aligned}$$

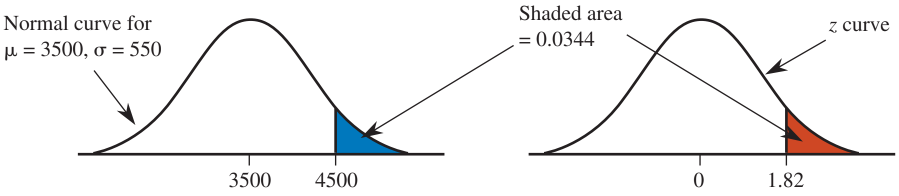


Figure 7.36 $P(x > 4500)$ and corresponding z curve area for the birth weight distribution of Example 7.30

Copyright 2025 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s). Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

Example 7.31 IQ Scores

Understand the context ➤ Although there is some controversy regarding the appropriateness of IQ scores as a measure of intelligence, they are still used for a variety of purposes. One commonly used IQ scale has a mean of 100 and a standard deviation of 15, and these IQ scores are approximately normally distributed. (Because this IQ score is based on the number of correct responses on a test, it is actually a discrete variable. However, the population distribution of IQ scores closely resembles a normal curve.) If you define the random variable

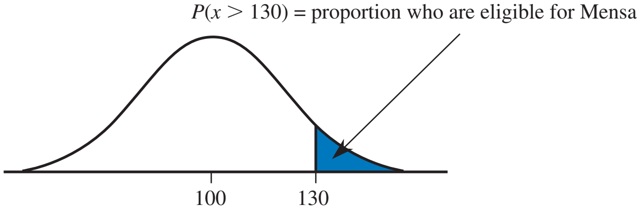
x = IQ score of a randomly selected individual

then x has a distribution that is approximately normal with $\mu = 100$ and $\sigma = 15$.

One way to become eligible for membership in Mensa, an organization for those of high intelligence, is to have an IQ score above 130. What proportion of the population would qualify for Mensa membership?

Do the work ➤ Answering this question requires evaluating $P(x > 130)$. This probability is shown in Figure 7.37.

Figure 7.37
Normal distribution and proportion greater than 130 for Example 7.31



With $a = 130$,

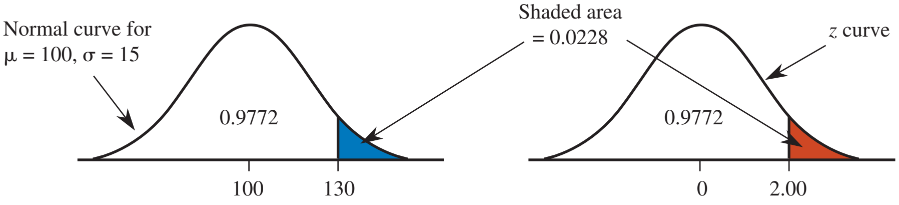
$$a^* = \frac{a - \mu}{\sigma} = \frac{130 - 100}{15} = 2.00$$

So (see Figure 7.38)

$$\begin{aligned} P(x > 130) &= P(z > 2.00) \\ &= z \text{ curve area to the right of } 2.00 \\ &= 1 - (z \text{ curve area to the left of } 2.00) \\ &= 1 - 0.9772 \\ &= 0.0228 \end{aligned}$$

Only about 2.28% of the population would qualify for Mensa membership.

Figure 7.38
$P(x > 130)$ and corresponding z curve for the IQ distribution of Example 7.31



Suppose that you are interested in the proportion of the population with IQ scores below 80—that is, $P(x < 80)$. With $b = 80$,

$$b^* = \frac{b - \mu}{\sigma} = \frac{80 - 100}{15} = -1.33$$

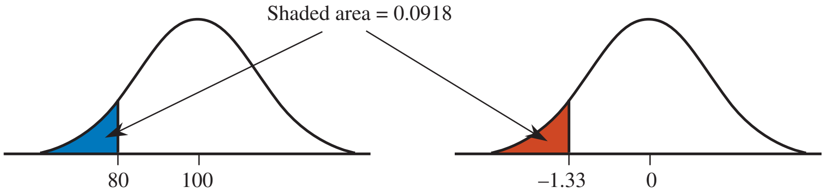
So

$$\begin{aligned} P(x < 80) &= P(z < -1.33) \\ &= z \text{ curve area to the left of } -1.33 \\ &= 0.0918 \end{aligned}$$

as shown in Figure 7.39. This probability tells you that just a little over 9% of the population has an IQ score below 80.

Copyright 2025 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s). Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

Figure 7.39
$P(x < 80)$ and corresponding z curve area for the IQ distribution of Example 7.31



Now consider the proportion of the population with IQs between 75 and 125. Using $a = 75$ and $b = 125$, you obtain

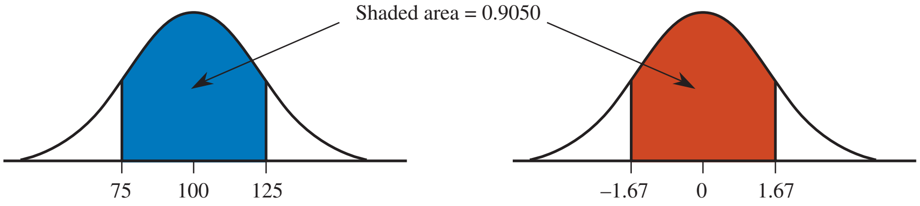
$$a^* = \frac{75 - 100}{15} = -1.67 \qquad b^* = \frac{125 - 100}{15} = 1.67$$

so

$$\begin{aligned} P(75 < x < 125) &= P(-1.67 < z < 1.67) \\ &= z \text{ curve area between } -1.67 \text{ and } 1.67 \\ &= (z \text{ curve area to the left of } 1.67) \\ &\quad - (z \text{ curve area to the left of } -1.67) \\ &= 0.9525 - 0.0475 \\ &= 0.9050 \end{aligned}$$

This is illustrated in Figure 7.40. This probability tells you that about 90.5% of the population has an IQ score between 75 and 125. Of the 9.5% with an IQ score that is not between 75 and 125, about half of them (4.75%) have scores greater than 125, and about half have scores less than 75.

Figure 7.40
$P(75 < x < 125)$ and corresponding z curve area



When translating a problem involving a normal distribution with mean μ and standard deviation σ to a problem involving the standard normal distribution, you convert to z-scores:

$$z = \frac{x - \mu}{\sigma}$$

Recall from Chapter 4 that a z-score can be interpreted as the distance of an x value from the mean in units of the standard deviation. A z-score of 1.4 corresponds to an x value that is 1.4 standard deviations above the mean, and a z-score of -2.1 corresponds to an x value that is 2.1 standard deviations below the mean.

Suppose that you are trying to evaluate $P(x < 60)$ for a variable whose distribution is normal with $\mu = 50$ and $\sigma = 5$. Converting the endpoint 60 to a z-score gives

$$z = \frac{60 - 50}{5} = 2$$

which tells you that the value 60 is two standard deviations above the mean. You then have

$$P(x < 60) = P(z < 2)$$

where z is a standard normal variable. Notice that for the standard normal distribution, which has $\mu = 0$ and $\sigma = 1$, the value 2 is two standard deviations above the mean. The value $z = 2$ is located the same distance (measured in standard deviations) from the mean of the standard normal distribution as is the value $x = 60$ from the mean in the normal distribution with $\mu = 50$ and $\sigma = 5$. This is why the translation using z-scores results in an equivalent problem involving the standard normal distribution.

Describing Extreme Values in a Normal Distribution

To describe the extreme values for a normal distribution with mean μ and standard deviation σ, first solve the corresponding problem for the standard normal distribution and then translate your answer back to the normal distribution of interest. This process is illustrated in Example 7.32.

Example 7.32 Registration Times

Understand the context ➤

Data on the length of time (in minutes) required to register for classes using an online system suggest that the distribution of the variable

x = time to register

for students at a particular university can be well approximated by a normal distribution with mean $\mu = 12$ minutes and standard deviation $\sigma = 2$ minutes. Because some students do not log off properly, the university would like to log off students automatically after some amount of time has elapsed. The university has decided to choose this time so that only 1% of the students are logged off while they are still attempting to register.

Do the work ➤

To determine the amount of time that should be allowed before disconnecting a student, you need to describe the largest 1% of the distribution of time to register. These are the individuals who will be mistakenly disconnected. This is illustrated in Figure 7.41. To determine the value of x^*, first solve the corresponding problem for the standard normal distribution.

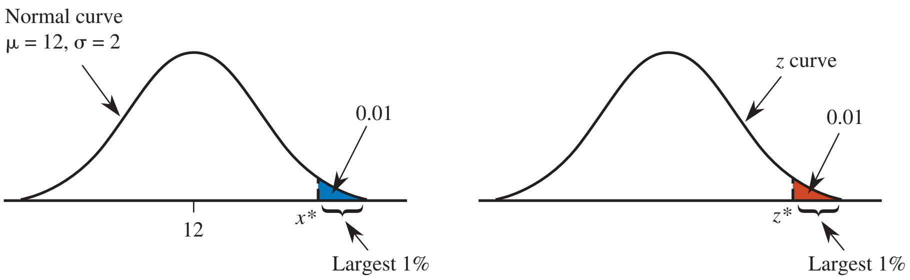


Figure 7.41
Capturing the largest 1% in a normal distribution

By looking for a cumulative area of 0.99 in Appendix Table 2, you find the closest entry (0.9901) in the 2.3 row and the .03 column, so $z^* = 2.33$. For the standard normal distribution, the largest 1% of the distribution is made up of those values greater than 2.33. This implies that in the distribution of time to register x (or any other normal distribution), the largest 1% are those values with z-scores greater than 2.33 or, equivalently, those x values more than 2.33 standard deviations above the mean. Here, the standard deviation is 2, so 2.33 standard deviations is 2.33(2), and it follows that

$$x^* = 12 + 2.33(2) = 12 + 4.66 = 16.66$$

Interpret the results ➤

The largest 1% of the distribution for time to register is made up of values that are greater than 16.66 minutes. If the system was set to log off students after 16.66 minutes, only about 1% of the students registering would be logged off before completing their registration.

A general formula for converting a z-score back to an x value results from solving

$$z^* = \frac{x^* - \mu}{\sigma}$$

for x^*, as shown in the accompanying box.

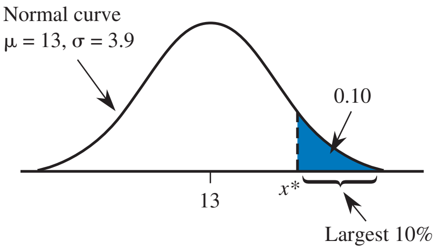
To convert a z-score, denoted by z^* back to an x value, use

$$x^* = \mu + z^*\sigma$$

Example 7.33 Garbage Truck Processing Times

Understand the context ➤

Garbage trucks entering a particular waste management facility are weighed before offloading garbage into a landfill. Data from the paper **"Estimating Waste Transfer Station Delays Using GPS" (*Waste Management* [2008]: 1742–1750)** suggest that a normal distribution with mean $\mu = 13$ minutes and $\sigma = 3.9$ minutes is a reasonable model for the probability distribution of the random variable x = total processing time for a garbage truck at this waste management facility (total processing time includes waiting time as well). Suppose that you want to describe the total processing times of the trucks making up the 10% with the longest processing times. These trucks would be the 10% with times corresponding to the shaded region in the accompanying figure.



Do the work ➤

For the standard normal distribution, the largest 10% are those with z-scores greater than $z^* = 1.28$ (from Appendix Table 2, based on a cumulative area of 0.90).

Then

$$\begin{aligned} x^* &= \mu + z^*\sigma \\ &= 13 + 1.28(3.9) \\ &= 13 + 4.992 \\ &= 17.992 \end{aligned}$$

About 10% of the garbage trucks using this facility would have a total processing time of more than about 18 minutes.

The 5% with the fastest processing times would be those with z-scores less than $z^* = -1.645$ (from Appendix Table 2, based on a cumulative area of 0.05). Then

$$\begin{aligned} x^* &= \mu + z^*\sigma \\ &= 13 + (-1.645)(3.9) \\ &= 13 - 6.416 \\ &= 6.584 \end{aligned}$$

About 5% of the garbage trucks processed at this facility will have total processing times of less than about 6.6 minutes.

Exercises 7.70–7.96

7.70 Determine the following standard normal (z) curve areas. (Hint: See Examples 7.23 and 7.24, or 7.26.)
- **a.** The area under the z curve to the left of 1.75
- **b.** The area under the z curve to the left of -0.68
- **c.** The area under the z curve to the right of 1.20
- **d.** The area under the z curve to the right of -2.82
- **e.** The area under the z curve between -2.22 and 0.53
- **f.** The area under the z curve between -1 and 1
- **g.** The area under the z curve between -4 and 4

7.71 Determine each of the following areas under the standard normal (z) curve:
- **a.** To the left of -1.28
- **b.** To the right of 1.28
- **c.** Between -1 and 2
- **d.** To the right of 0
- **e.** To the right of -5
- **f.** Between -1.6 and 2.5
- **g.** To the left of 0.23

7.72 Let z denote a random variable that has a standard normal distribution. Determine each of the following probabilities:

a. $P(z < 2.36)$
b. $P(z \leq 2.36)$
c. $P(z < -1.23)$
d. $P(1.14 < z < 3.35)$

7.73 Let z denote a random variable that has a standard normal distribution. Determine each of the following probabilities:

a. $P(-0.77 \leq z \leq -0.55)$
b. $P(z > 2)$
c. $P(z \geq -3.38)$
d. $P(z < 4.98)$

7.74 Let z denote a variable that has a standard normal distribution. Determine the value z^* to satisfy the following conditions. (Hint: See Example 7.24 or 7.28.)

a. $P(z < z^*) = 0.025$
b. $P(z < z^*) = 0.01$
c. $P(z < z^*) = 0.05$
d. $P(z > z^*) = 0.02$
e. $P(z > z^*) = 0.01$
f. $P(z > z^* \text{ or } z < -z^*) = 0.20$

7.75 Determine the value z^* that

a. Separates the largest 3% of all z values from the others
b. Separates the largest 1% of all z values from the others
c. Separates the smallest 4% of all z values from the others
d. Separates the smallest 10% of all z values from the others

7.76 Determine the value of z^* such that

a. $-z^*$ and z^* separate the middle 95% of all z values from the most extreme 5%
b. $-z^*$ and z^* separate the middle 90% of all z values from the most extreme 10%
c. $-z^*$ and z^* separate the middle 98% of all z values from the most extreme 2%
d. $-z^*$ and z^* separate the middle 92% of all z values from the most extreme 8%

7.77 Because $P(z < 0.44) = 0.67$, 67% of all z values are less than 0.44, and 0.44 is the 67th percentile of the standard normal distribution. Determine the value of each of the following percentiles for the standard normal distribution. (Hint: If the cumulative area that you must look for does not appear in the z table, use the closest entry.)

a. The 91st percentile (Hint: Look for area 0.9100.)
b. The 77th percentile
c. The 50th percentile
d. The 9th percentile
e. What is the relationship between the 70th z percentile and the 30th z percentile?

7.78 Consider the population of all 1-gallon cans of dusty rose paint manufactured by a particular paint company. Suppose that a normal distribution with mean $\mu = 5$ ml and standard deviation $\sigma = 0.2$ ml is a reasonable model for the distribution of the variable $x =$ amount of red dye in the paint mixture. Use the normal distribution model to calculate the following probabilities.

a. $P(x < 5.0)$
b. $P(x < 5.4)$
c. $P(x \leq 5.4)$
d. $P(4.6 < x < 5.2)$
e. $P(x > 4.5)$
f. $P(x > 4.0)$

7.79 The number of vehicles leaving a highway at a certain exit during a particular time period has a distribution that is approximately normal with mean value 500 and standard deviation 75. What is the probability that the number of cars exiting during this period is

a. At least 650?
b. Strictly between 400 and 550? (*Strictly* means that the values 400 and 550 are not included.)
c. Between 400 and 550 (inclusive)?

7.80 A pizza company advertises that it puts 0.5 pound of real mozzarella cheese on its medium-sized pizzas. In fact, the amount of cheese on a randomly selected medium pizza is normally distributed with a mean value of 0.5 pound and a standard deviation of 0.025 pound.

a. What is the probability that the amount of cheese on a medium pizza is between 0.525 and 0.550 pounds?
b. What is the probability that the amount of cheese on a medium pizza exceeds the mean value by more than 2 standard deviations?
c. What is the probability that three randomly selected medium pizzas each have at least 0.475 pounds of cheese?

7.81 A machine that produces ball bearings has initially been set so that the mean diameter of the bearings it produces is 0.500 inches. A bearing is acceptable if its diameter is within 0.004 inches of this target value. Suppose, however, that the setting has changed during the course of production, so that the distribution of the diameters produced is now approximately normal with mean 0.499 inch and standard deviation 0.002 inch. What percentage of the bearings produced will not be acceptable?

7.82 The report **"The Common Sense Census: Media Use by Tweens and Teens" (Common Sense Media, 2019)** used data from a large representative sample of 13- to 18-year-olds who report that they send texts to construct the following table.

Copyright 2025 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s). Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

Number of Texts Sent per Day	Relative Frequency
1–10	0.40
11–30	0.34
31–50	0.12
More than 50	0.14

Also reported was that the average number of texts sent per day for this sample was 39. Explain why it is not reasonable to think that the distribution of x = Number of texts sent per day is approximately normal for the population of 13- to 18-year-olds.

7.83 Consider babies born in the "normal" range of 37–43 weeks gestational age. The paper referenced in Example 7.30 **"Birth Weight Curves Tailored to Maternal World Region" (*Journal of Obstetrics and Gynaecology Canada* [2012]: 159–171)** suggests that a normal distribution with mean $\mu = 3500$ grams and standard deviation $\sigma = 600$ grams is a reasonable model for the probability distribution of the variable x = birth weight of a randomly selected full-term baby born in Canada.

a. What is the probability that the birth weight of a randomly selected full-term baby born in Canada exceeds 4000 g?

b. What is the probability that the birth weight of a randomly selected full-term baby born in Canada is between 3000 and 4000 g?

c. What is the probability that the birth weight of a randomly selected full-term baby born in Canada is either less than 2000 g or greater than 5000 g?

7.84 Use the information on birth weights for babies born in Canada given in the previous exercise to answer the following questions.

a. What is the probability that the birth weight of a randomly selected full-term baby born in Canada exceeds 7 pounds? (Hint: 1 lb = 453.59 g.)

b. How would you characterize the most extreme 0.1% of all full-term baby birth weights for babies born in Canada?

c. If x is a random variable with a normal distribution and a is a numerical constant ($a \neq 0$), then $y = ax$ also has a normal distribution. Use this formula to determine the distribution of full-term baby birth weight expressed in pounds (shape, mean, and standard deviation), and then recalculate the probability from Part (a). How does this compare to your previous answer?

7.85 Emissions of nitrogen oxides, which are major constituents of smog, can be modeled using a normal distribution. Let x denote the amount of this pollutant emitted (in parts per billion) by a randomly selected vehicle. Suppose the distribution of x can be described by a normal distribution with $\mu = 1.6$ and $\sigma = 0.4$. A city wants to offer some sort of incentive to get the worst polluters off the road. What emission levels constitute the worst 10% of the vehicles?

7.86 The paper referenced in Example 7.33 **("Estimating Waste Transfer Station Delays Using GPS," *Waste Management* [2008]: 1742–1750)** describing processing times for garbage trucks also provided information on processing times at a second facility. At this second facility, the mean total processing time was 9.9 minutes and the standard deviation of the processing times was 6.2 minutes. Explain why a normal distribution with mean 9.9 and standard deviation 6.2 would not be an appropriate model for the probability distribution of the variable x = total processing time of a randomly selected truck entering this facility.

7.87 The size of the left upper chamber of the heart is one measure of cardiovascular health. When the upper left chamber is enlarged, the risk of heart problems is increased. The paper **"Left Atrial Size Increases with Body Mass Index in Children" (*International Journal of Cardiology* [2009]: 1–7)** described a study in which the left atrial size was measured for a large number of children age 5 to 15 years. Based on these data, the authors concluded that for healthy children, left atrial diameter was approximately normally distributed with a mean of 26.4 mm and a standard deviation of 4.2 mm.

a. Approximately what proportion of healthy children have left atrial diameters less than 24 mm?

b. Approximately what proportion of healthy children have left atrial diameters greater than 32 mm?

c. Approximately what proportion of healthy children have left atrial diameters between 25 and 30 mm?

d. For healthy children, what is the value for which only about 20% have a larger left atrial diameter?

7.88 The paper referenced in the previous exercise also included data on left atrial diameter for children who were considered overweight. For these children, left atrial diameter was approximately normally distributed with a mean of 28 mm and a standard deviation of 4.7 mm.

a. Approximately what proportion of overweight children have left atrial diameters less than 25 mm?

b. Approximately what proportion of overweight children have left atrial diameters greater than 32 mm?

c. Approximately what proportion of overweight children have left atrial diameters between 25 and 30 mm?

d. What proportion of overweight children has left atrial diameters greater than the mean for healthy children?

7.89 The article **"New York City's Graffiti-Removal Response Time Rises" (*The Wall Street Journal*, September 16, 2016)** states that the city took an average of 114 days to handle graffiti complaints in 2015. Suppose that the response time is approximately normally distributed with a mean of 114 days and a standard deviation of 20 days.

a. Approximately what proportion of graffiti removal requests are handled within 60 days?
b. Approximately what proportion of graffiti removal requests take more than 120 days?

7.90 A machine that cuts corks for wine bottles operates in such a way that the distribution of the diameter for the corks produced is well approximated by a normal distribution with mean 3 cm and standard deviation 0.1 cm. The specifications call for corks with diameters between 2.9 and 3.1 cm. A cork not meeting the specifications is considered defective. (A cork that is too small leaks and causes the wine to deteriorate. A cork that is too large doesn't fit in the bottle.) What proportion of corks produced by this machine are defective?

7.91 Refer to the previous exercise. Suppose that there are two machines available for cutting corks. The machine described in the preceding problem produces corks with diameters that are approximately normally distributed with mean 3 cm and standard deviation 0.1 cm. The second machine produces corks with diameters that are approximately normally distributed with mean 3.05 cm and standard deviation 0.01 cm. Which machine would you recommend? (Hint: Which machine would produce fewer defective corks?)

7.92 Purchases made at small "corner stores" were studied by the authors of the paper **"Changes in Quantity, Spending, and Nutritional Characteristics of Adult, Adolescent and Child Urban Corner Store Purchases After an Environmental Intervention" (*Preventive Medicine* [2015]: 81–85)**. Corner stores were defined as stores that are less than 200 square feet in size, have only one cash register, and primarily sell food. After observing a large number of corner store purchases in Philadelphia, the authors reported that the average number of grams of fat in a corner store purchase was 21.1. Suppose that the variable x = number of grams of fat in a corner store purchase has a distribution that is approximately normal with a mean of 21.1 grams and a standard deviation of 7 grams.

a. What is the probability that a randomly selected corner store purchase has more than 30 grams of fat?
b. What is the probability that a randomly selected corner store purchase has between 15 and 25 grams of fat?
c. If two corner store purchases are randomly selected, what is the probability that both of these purchases will have more than 25 grams of fat?

7.93 Suppose that fuel efficiency (miles per gallon, mpg) for a particular car model under specified conditions is normally distributed with a mean value of 30.0 mpg and a standard deviation of 1.2 mpg.

a. What is the probability that the fuel efficiency for a randomly selected car of this model is between 29 and 31 mpg?
b. Would it surprise you to find that the efficiency of a randomly selected car of this model is less than 25 mpg?
c. If three cars of this model are randomly selected, what is the probability that each of the three have efficiencies exceeding 32 mpg?
d. Find a number x^* such that 95% of all cars of this model have efficiencies exceeding x^* (i.e., $P(x > x^*) = 0.95$).

7.94 Suppose that the distribution of typing speed in words per minute (wpm) for experienced typists using a new type of split keyboard can be approximated by a normal curve with mean 60 wpm and standard deviation 15 wpm **("The Effects of Split Keyboard Geometry on Upper body Postures," *Ergonomics* [2009]: 104–111)**.

a. What is the probability that a randomly selected typist's speed is at most 60 wpm?
b. What is the probability that a randomly selected typist's speed is less than 60 wpm?
c. What is the probability that a randomly selected typist's speed is between 45 and 90 wpm?

7.95 When used in a particular digital camera, the lifetime of a certain brand of battery is normally distributed with a mean value of 6 hours and a standard deviation of 0.8 hour. Suppose that two new batteries are independently selected and put into the camera. The camera stops working as soon as one of the two batteries fails.

a. What is the probability that the digital camera works for at least 4 hours?
b. What is the probability that the digital camera works for at most 7 hours?
c. Find a number x^* such that only 5% of all digital cameras will work without battery replacement for more than x^* hours.

7.96 A machine producing vitamin E capsules operates so that the actual amount of vitamin E in each capsule is normally distributed with a mean of 5 mg and a standard deviation of 0.05 mg. What is the probability that a randomly selected capsule contains less than 4.9 mg of vitamin E? At least 5.2 mg of vitamin E?

Section 7.7 Checking for Normality and Normalizing Transformations

Some of the most frequently used statistical methods are valid only when the sample has been selected from a population distribution that is at least approximately normal. One way to determine if it is reasonable to assume that the population distribution is approximately normal is to construct a **normal probability plot** of the sample data.

A normal probability plot uses quantities called **normal scores**. The values of the normal scores depend on the sample size n. For example, the normal scores when $n = 10$ are as follows:

−1.54	−1.00	−0.66	−0.38	−0.12
0.12	0.38	0.66	1.00	1.54

To interpret these numbers, think of selecting sample after sample from a standard normal distribution, each one consisting of $n = 10$ observations. Then −1.54 is the long-run average of the smallest observation from each sample, −1.00 is the long-run average of the second smallest observation from each sample, and so on.

Tables of normal scores for many different sample sizes are available. Alternatively, many statistical software packages (such as Minitab and JMP) and some graphing calculators can compute these scores on request and then use them to construct a normal probability plot. Not all calculators and software packages use the same algorithm to compute normal scores. However, these differences in normal scores do not change the overall character of a normal probability plot, so either values from a table or those provided by software or a calculator can be used.

After the sample observations are ordered from smallest to largest, the smallest normal score is paired with the smallest observation in the sample, the second smallest normal score with the second smallest observation in the sample, and so on. The first number in a pair is the normal score, and the second number in the pair is the observed data value. A normal probability plot is a scatterplot of these (normal score, observed value) pairs.

If the sample has been selected from a population with a *standard* normal distribution, the second number in each pair should be reasonably close to the first number (ordered observation ≈ corresponding normal score). Then the n plotted points will fall near a line with slope equal to 1 (a 45° line) passing through (0, 0). When the sample has been selected from *some* normal population distribution (but not necessarily the standard normal distribution), the plotted points will be close to *some* straight line (but not necessarily one with slope 1 and intercept 0).

Definition

Normal probability plot: A scatterplot of the (normal score, observed value) pairs.

A strong linear pattern in a normal probability plot suggests it is reasonable to think that the population distribution is normal. On the other hand, systematic departure from a straight-line pattern (such as curvature in the plot) indicates that it is not reasonable to assume that the population distribution is normal.

Example 7.34 Egg Weights

Understand the context ➤

● The following data represent egg weights (in grams) for a sample of 10 eggs. These data are consistent with summary quantities in the paper **"Evaluation of Egg Quality Traits of Chickens Reared under Backyard System in Western Uttar Pradesh" (*Indian Journal of Poultry Science,* 2009).**

● **Data set available online**

53.04	53.50	52.53	53.00	53.07	52.86	52.66	53.23	53.26	53.16

Copyright 2025 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s). Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

Do the work ➤

Arranging the sample observations in order from smallest to largest results in

52.53	52.66	52.86	53.00	53.04	53.07	53.16	53.23	53.26	53.50

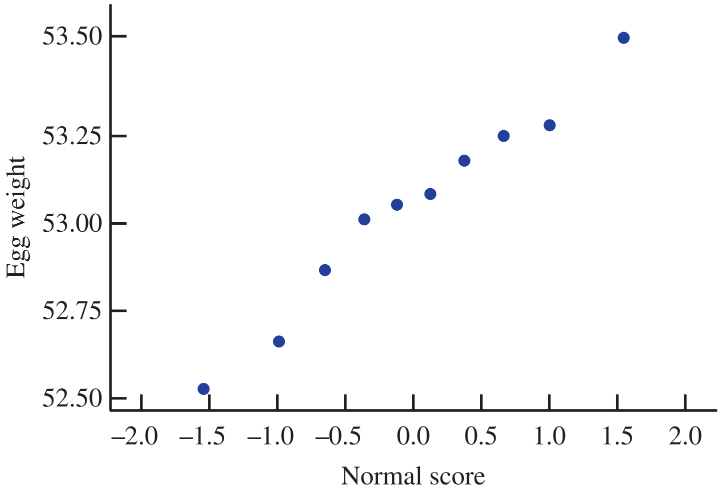
Pairing these ordered observations with the normal scores for a sample of size 10 (previously given) results in the following 10 pairs that can be used to construct the normal probability plot:

(−1.54, 52.53)	(−1.00, 52.66)
(−0.66, 52.86)	(−0.38, 53.00)
(−0.12, 53.04)	(0.12, 53.07)
(0.38, 53.16)	(0.66, 53.23)
(1.00, 53.26)	(1.54, 53.50)

Interpret the results ➤

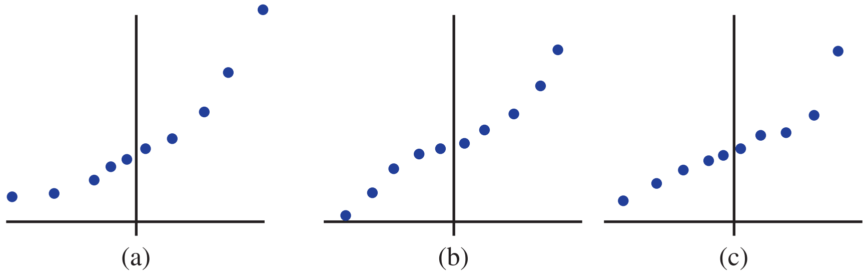
The normal probability plot is shown in Figure 7.42. The linear pattern in the plot suggests that it is reasonable to think that the egg-weight distribution from which these observations were selected is approximately normal.

Figure 7.42
A normal probability plot for data of Example 7.34.



Deciding whether a normal probability plot shows a strong linear pattern is sometimes a matter of opinion. Particularly when n is small, normality should not be ruled out unless the pattern is clearly nonlinear. Figure 7.43 displays several plots that suggest that the population distribution may not be normal.

Figure 7.43
Plots suggesting nonnormality:
(a) indication that the population distribution is skewed;
(b) indication that the population distribution has heavier tails than a normal curve;
(c) presence of an outlier.



Transforming Data to Obtain a Distribution That Is Approximately Normal (Optional)

Many statistical methods are valid only when the sample is selected at random from a population whose distribution is at least approximately normal. When a sample histogram shows a distinctly nonnormal shape, it is common to use a transformation or reexpression of the data.

Copyright 2025 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s). Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

When *transforming* data, a specified mathematical function (such as the square root, logarithm, or reciprocal) is applied to each data value to produce a set of transformed data. You then study and summarize the distribution of these transformed values.

You saw in Chapter 5 that, with bivariate data, one or both of the variables can be transformed in an attempt to find two variables that are linearly related. With univariate data, a transformation is usually chosen to produce a distribution of transformed values that is more nearly symmetric and more closely approximated by a normal distribution than the original distribution.

Example 7.35 Rainfall Data

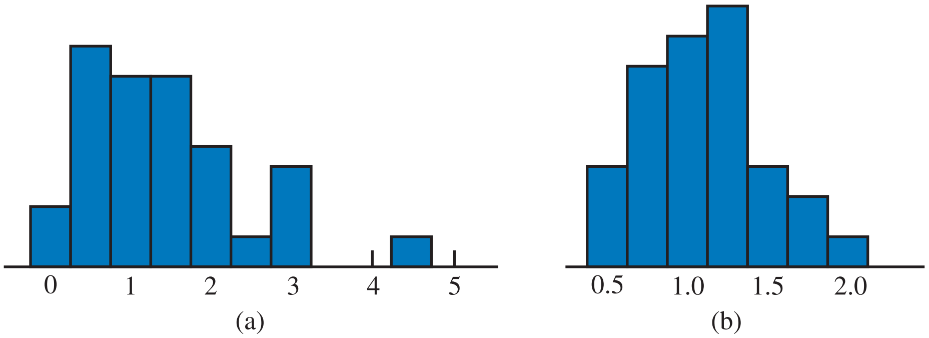
Understand the context ➤

● Data that have been used by several investigators to introduce the concept of transformation consist of values of March precipitation for Minneapolis–St. Paul over a period of 30 years. These values are given in Table 7.2, along with the square root of each value. Histograms of both the original and the transformed data appear in Figure 7.44.

Table 7.2 Original and Square-Root-Transformed Values of March Precipitation in Minneapolis–St. Paul over a 30-year Period

Year	Precipitation	$\sqrt{\text{Precipitation}}$	Year	Precipitation	$\sqrt{\text{Precipitation}}$
1	0.77	0.88	16	1.62	1.27
2	1.74	1.32	17	1.31	1.14
3	0.81	0.90	18	0.32	0.57
4	1.20	1.10	19	0.59	0.77
5	1.95	1.40	20	0.81	0.90
6	1.20	1.10	21	2.81	1.68
7	0.47	0.69	22	1.87	1.37
8	1.43	1.20	23	1.18	1.09
9	3.37	1.84	24	1.35	1.16
10	2.20	1.48	25	4.75	2.18
11	3.00	1.73	26	2.48	1.57
12	3.09	1.76	27	0.96	0.98
13	1.51	1.23	28	1.89	1.37
14	2.10	1.45	29	0.90	0.95
15	0.52	0.72	30	2.05	1.43

Figure 7.44
Histograms of the precipitation data used in Example 7.35:
(a) untransformed data;
(b) square-root transformed data.



The distribution of the original data is clearly skewed, with a long upper tail. The square-root transformation results in a distribution that is more symmetric, with a typical value of around 1.1.

● **Data set available online**

Logarithmic transformations are also common and, as with bivariate data, either the natural logarithm or the base 10 logarithm can be used. A logarithmic transformation is often used with data that are positively skewed (a long upper tail). This affects values in the upper tail substantially more than values in the lower tail, resulting in a more symmetric—and often more nearly normal—distribution.

Example 7.36 Markers for Kidney Disease

Understand the context ➤

● Two measures of kidney function are the levels of a substance called AGT found in blood and in urine. The paper **"Urinary Angiotensinogen as a Potential Biomarker of Severity of Chronic Kidney Diseases" (*Journal of the American Society of Hypertension* [2008]: 349–354)** describes a study in which blood plasma AGT levels and urinary AGT levels were measured for a sample of adults with chronic kidney disease. Representative data (consistent with summary quantities and descriptions given in the paper) for 40 patients are given in Table 7.3.

Table 7.3 Plasma and Urinary AGT Levels

Plasma AGT	Plasma AGT	Urinary AGT	Urinary AGT
21.0	16.7	56.2	41.7
36.0	20.2	288.4	29.5
22.9	24.5	45.7	208.9
8.0	18.5	426.6	229.1
27.3	40.2	190.6	186.2
32.4	18.8	616.6	29.5
17.2	28.1	97.7	229.1
30.9	26.8	66.1	13.5
27.2	24.1	2.6	407.4
30.0	14.1	74.1	1122.0
35.1	18.9	14.5	66.1
21.6	25.6	56.2	7.4
22.7	10.2	812.8	177.8
2.5	29.2	11.5	6.2
30.2	29.5	346.7	67.6
27.3	24.3	9.6	20.0
19.6	22.3	288.4	28.8
19.0	16.5	147.9	186.2
13.4	25.6	17.0	141.3
18.0	23.0	575.4	724.4

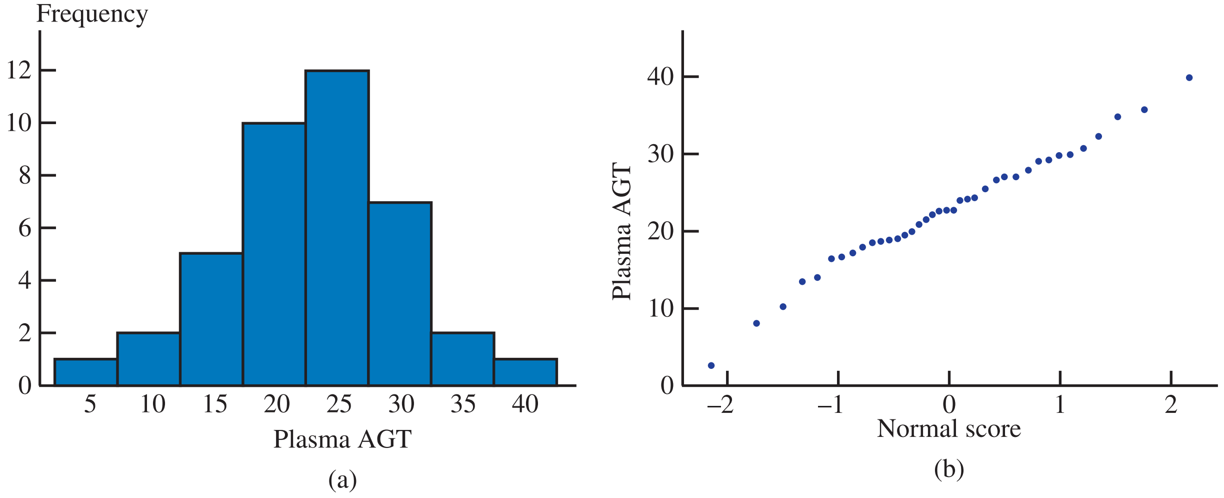
The authors of the paper stated that the distribution of plasma AGT levels was approximately normal. Minitab was used to construct the histogram and normal probability plot for the plasma AGT levels shown in Figure 7.45. The histogram is reasonably symmetric and the normal probability plot shows a strong linear pattern. This is consistent with the authors' statement the approximate normality of the plasma AGT levels.

When the authors considered urinary AGT levels, they found that the distribution of the sample data was skewed, and they used a log transformation in order to obtain a distribution that was more approximately normal. Table 7.4 gives the urinary AGT

● **Data set available online**

Figure 7.45

Graphical displays for the plasma AGT data of Example 7.36: (a) histogram; (b) normal probability plot.



levels along with the log-transformed data. Figure 7.46 shows histograms of the original urinary AGT data and the transformed urinary AGT data. Notice that the histogram for the transformed data is more nearly symmetric and more mound shaped than the histogram of the untransformed data.

Table 7.4 Urinary AGT Levels and Log-Transformed Levels

Urinary AGT	Log(Urinary AGT)	Urinary AGT	Log(Urinary AGT)
56.2	1.75	41.7	1.62
288.4	2.46	29.5	1.47
45.7	1.66	208.9	2.32
426.6	2.63	229.1	2.36
190.6	2.28	186.2	2.27
616.6	2.79	29.5	1.47
97.7	1.99	229.1	2.36
66.1	1.82	13.5	1.13
2.6	0.41	407.4	2.61
74.1	1.87	1122.0	3.05
14.5	1.16	66.1	1.82
56.2	1.75	7.4	0.87
812.8	2.91	177.8	2.25
11.5	1.06	6.2	0.79
346.7	2.54	67.6	1.83
9.6	0.98	20.0	1.30
288.4	2.46	28.8	1.46
147.9	2.17	186.2	2.27
17.0	1.23	141.3	2.15
575.4	2.76	724.4	2.86

Copyright 2025 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s). Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

Figure 7.46
Histograms of urinary AGT data from Example 7.36: (a) untransformed data; (b) transformed data.

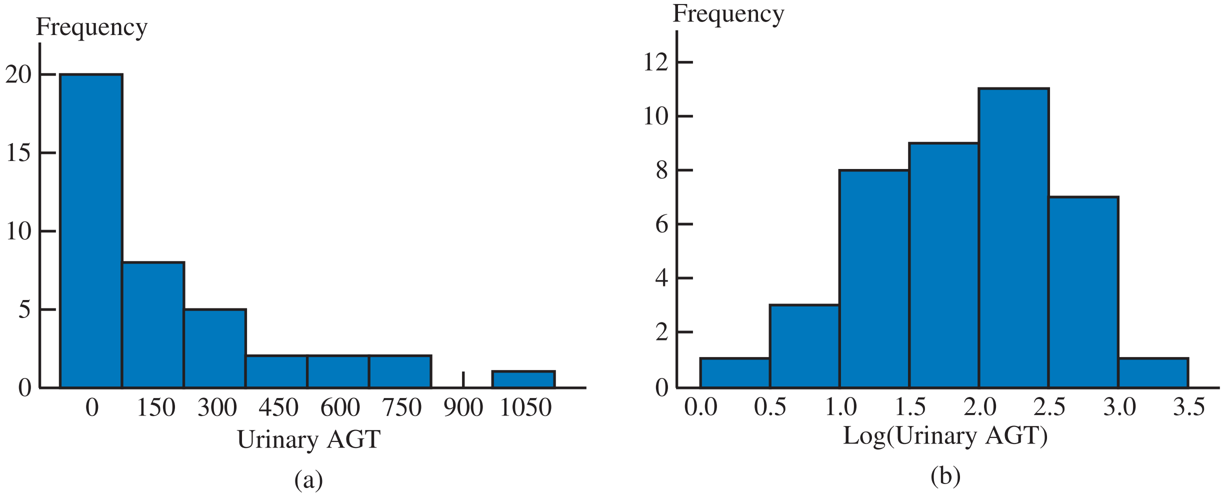
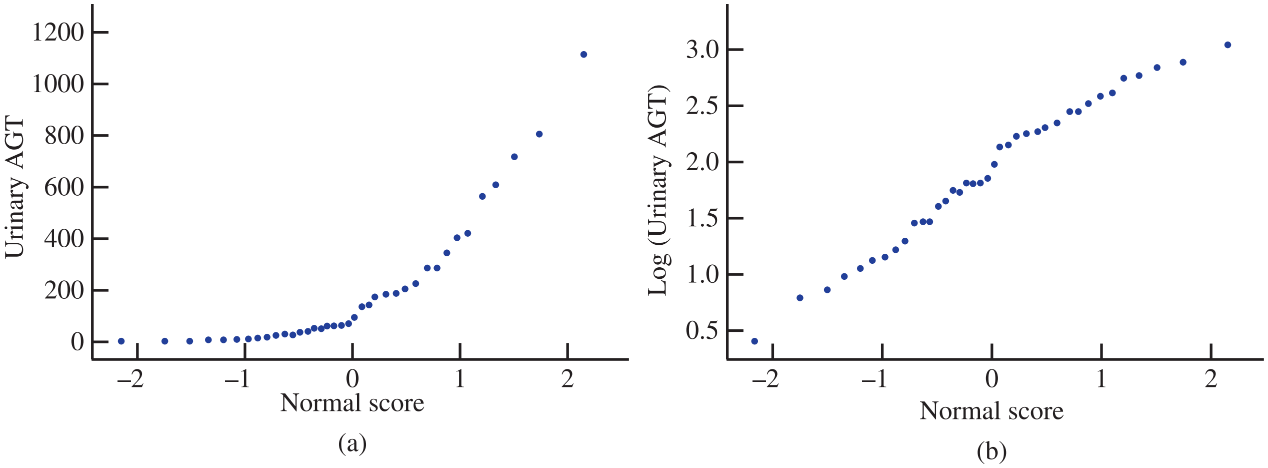


Figure 7.47 displays Minitab normal probability plots for the original data and for the transformed data. The plot for the transformed data is clearly more linear in appearance than the plot for the original data.

Figure 7.47
Minitab normal probability plots for the urinary AGT data: (a) original data; (b) transformed data.

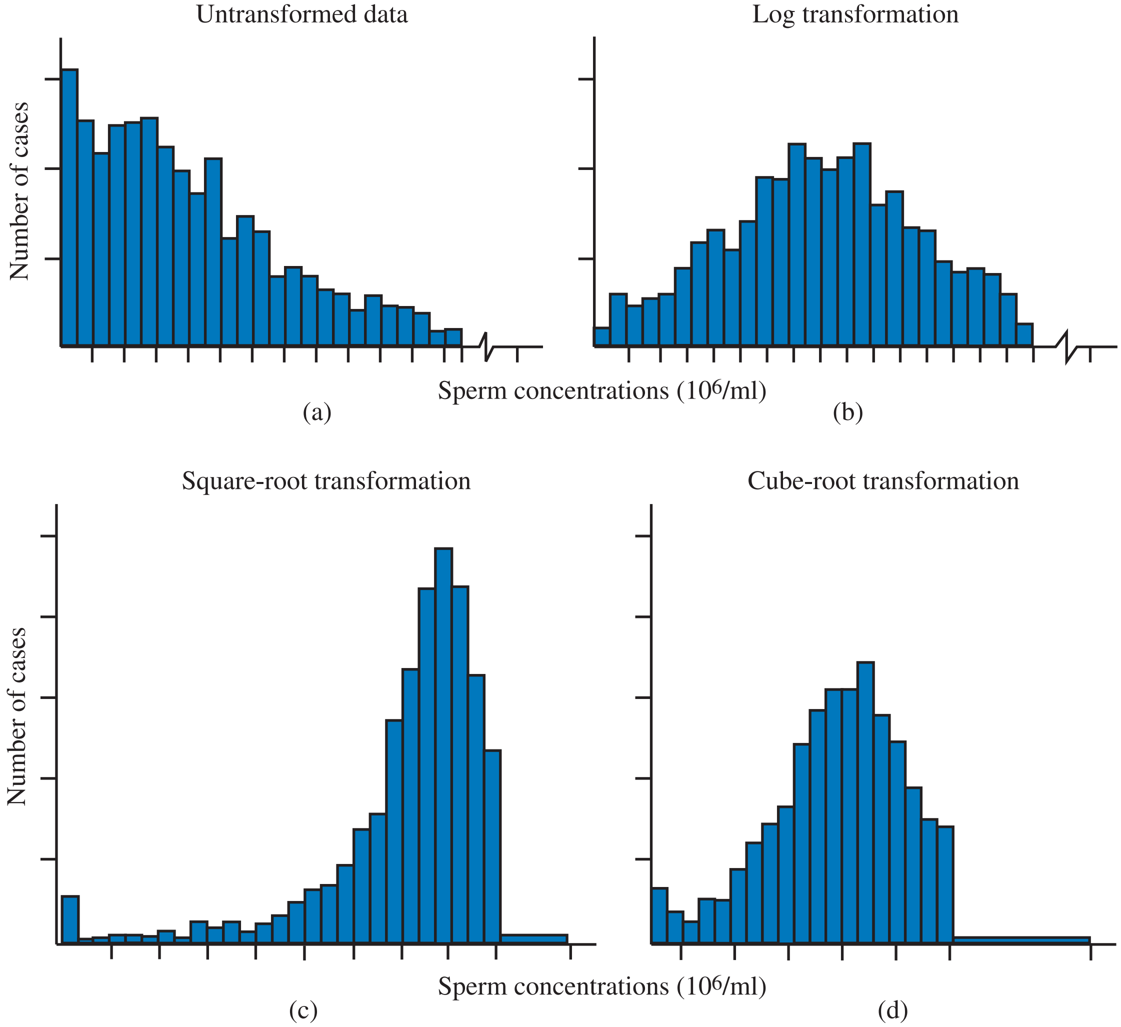


Selecting a Transformation

When selecting a transformation, it is common to try several different transformations to find one that is satisfactory. Figure 7.48, from the article **"Distribution of Sperm Counts in Suspected Infertile Men" (*Journal of Reproduction and Fertility* [1983]: 91–96)**, shows what can result from such a search. Investigators in this field have used all three of the transformations illustrated, but the log transformation shown in Figure 7.48(b) appears to be the most successful in creating a distribution that is approximately normal.

Figure 7.48

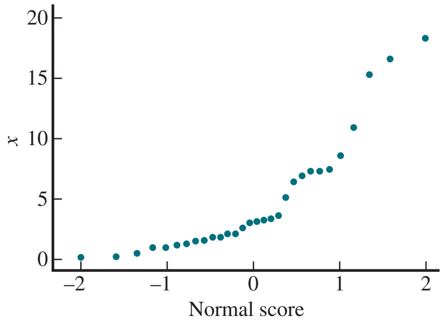
Histograms of sperm concentrations for 1711 suspected infertile men: (a) untransformed data (highly skewed); (b) log-transformed data (reasonably symmetric); (c) square-root-transformed data; (d) cube-root-transformed data.

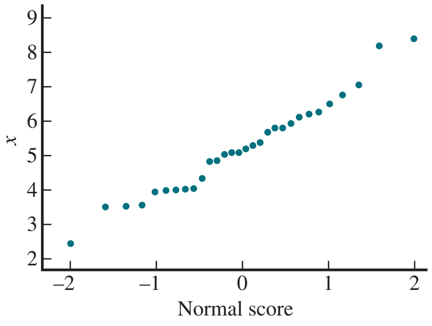


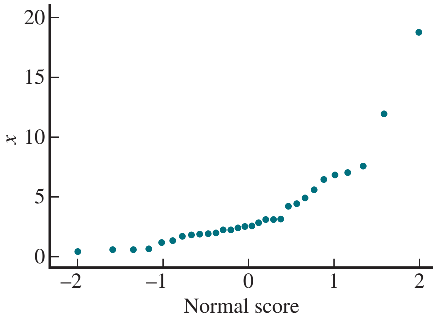
Exercises 7.97–7.108

● **Data set available online**

7.97 For each of the normal probability plots shown here, indicate whether you think it is reasonable to think that the population distribution is approximately normal and explain your thinking.







Copyright 2025 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s). Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

7.98 ● The authors of the paper **“Development of Nutritionally At-Risk Young Children Is Predicted by Malaria, Anemia, and Stunting in Pemba, Zanzibar” (*Journal of Nutrition* [2009]: 763–772)** studied factors that might be related to dietary deficiencies in children. Children were observed for a fixed period of time, and the amount of time spent in various activities was recorded. One variable of interest was the amount of time (in minutes) a child spent fussing. Data for 15 children, consistent with summary quantities in the paper, are given in the accompanying table. Normal scores for a sample size of 15 are also given.

Fussing Time (x)	Normal Score
0.05	−1.739
0.10	−1.245
0.15	−0.946
0.40	−0.714
0.70	−0.515
1.05	−0.335
1.95	−0.165
2.15	0.000
3.70	0.165
3.90	0.335
4.50	0.515
6.00	0.714
8.00	0.946
11.00	1.245
14.00	1.739

Construct a normal probability plot for the fussing time data. Does the plot look linear? Do you agree with the authors of the paper that the fussing time distribution is not normal? (Hint: See Example 7.34.)

7.99 ● Measures of nerve conductivity are used in the diagnosis of certain medical conditions. The paper **“Effects of Age, Gender, Height, and Weight on Late Responses and Nerve Conduction Study Parameters” (*Acta Neurologica Taiwanica* [2009]: 242–249)** describes a study in which the ulnar nerve was stimulated in healthy patients and the amplitude and velocity of the response was measured.

Representative data (consistent with summary quantities and descriptions given in the paper) for 30 patients for the variable x = response velocity (m/s) are given in the accompanying table. Also given are values of the log of x and the square root of x.

x	log(x)	sqrt(x)
60.1	1.78	7.75
48.7	1.69	6.98
51.7	1.71	7.19
52.9	1.72	7.27
50.5	1.70	7.11
58.5	1.77	7.65
53.6	1.73	7.32
60.3	1.78	7.77
64.5	1.81	8.03
50.4	1.70	7.10
56.5	1.75	7.52
55.5	1.74	7.45
53.0	1.72	7.28
50.5	1.70	7.11
54.0	1.73	7.35
53.6	1.73	7.32
55.2	1.74	7.43
57.9	1.76	7.61
61.5	1.79	7.84
58.0	1.76	7.62
57.6	1.76	7.59
67.1	1.83	8.19
56.2	1.75	7.50
53.8	1.73	7.33
55.7	1.75	7.46
52.9	1.72	7.27
54.0	1.73	7.35
52.6	1.72	7.25
61.8	1.79	7.86
62.8	1.80	7.92

a. Construct a histogram of the untransformed data.
b. Does the distribution of x appear to be approximately normal? Explain.
c. Construct a histogram of the log-transformed data.
d. Is the histogram of the log-transformed data more nearly symmetric than the histogram of the untransformed data?

7.100 Use the information given in the previous question to answer the following questions.
a. Construct a histogram of the square root transformed data.
b. Do either of the two transformations (square root or log) result in a histogram that is more nearly normal in shape? (Hint: See Example 7.36.)

7.101 ● Macular degeneration is the most common cause of blindness in people older than 60 years. One variable thought to be related to a type of inflammation associated with this disease is level of a substance called soluble Fas ligand (sFasL) in the blood.

The accompanying table contains representative data on x = sFasL level for 10 patients with age-related macular degeneration. These data are consistent with summary quantities and

Copyright 2025 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s). Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

descriptions of the data given in the paper **"Associations of Plasma-Soluble Fas Ligand with Aging and Age-Related Macular Degeneration" (*Investigative Ophthalmology* & *Visual Science* [2008]: 1345–1349)**.

The authors of the paper noted that the distribution of sFasL level was skewed and recommended a cube-root transformation. The cube-root values and the normal scores for a sample size of 10 are also given in the accompanying table.

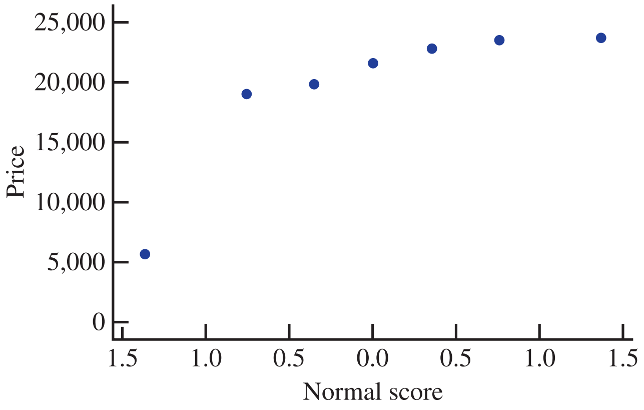
x	Cube Root of *x*	Normal Score
0.069	0.41	−1.539
0.074	0.42	−1.001
0.176	0.56	−0.656
0.185	0.57	−0.376
0.216	0.60	−0.123
0.287	0.66	0.123
0.343	0.70	0.376
0.343	0.70	0.656
0.512	0.80	1.001
0.729	0.90	1.539

a. Construct a normal probability plot using the untransformed data.

b. Does the normal probability plot for the untransformed data appear linear or curved?

c. Construct a normal probability plot using the cube-root transformed data. Does the normal probability plot appear more nearly linear than the plot for the untransformed data?

7.102 The following normal probability plot was constructed using data on the price of seven 2015 Honda Accords with automatic transmissions that were listed for sale within 25 miles of the zip code 19383 **(from autotrader.com, search conducted on September 24, 2016)**. For purposes of this exercise, you may assume that this sample is representative of 2015 Honda Accord prices in this area. Based on the normal probability plot, is it reasonable to think that the distribution of 2015 Honda Accord prices in this area is approximately normal? Explain.



7.103 ● Consider the following 10 observations on the lifetime (in hours) for a certain type of power supply: 152.7, 172.0, 172.5, 173.3, 193.0, 204.7, 216.5, 234.9, 262.6, and 422.6. Construct a normal probability plot, and comment on the plausibility of a normal distribution as a model for power supply lifetime. (The normal scores for a sample of size 10 are −1.539, −1.001, −0.656, −0.376, −0.123, 0.123, 0.376, 0.656, 1.001, and 1.539.)

7.104 ● The paper **"Investigating the Relation Among Disturbed Sleep Due to Social Media Use, School Burnout, and Academic Performance" (*Journal of Adolescence* [2020]: 156–164)** carried out a study in which they surveyed adolescents about their sleep patterns. Based on the data collected, they created a measure of sleep disturbance, with higher values representing greater disturbance. Sleep disturbance scores for a sample of 15 adolescents that are consistent with summary quantities given in the paper follow.

1.16	1.03	0.96	1.23	0.96	1.07	1.30	1.21
0.94	2.00	1.19	1.23	0.94	1.39	1.23	

a. Use these data to construct a normal probability plot (the normal scores for a sample of size 15 appear in Exercise 7.98).

b. The authors of the paper concluded that the distribution of sleep disturbance scores was not normal. Does the normal probability plot from Part (a) support this conclusion? Explain.

7.105 ● Example 7.35 examined rainfall data for Minneapolis–St. Paul. The square-root transformation was used to obtain a distribution of values that was more nearly symmetric than the distribution of the original data. Another transformation that has been suggested by meteorologists is the cube root: transformed value = (original value)$^{1/3}$. The original values ordered from smallest to largest and their cube roots (the transformed values) are given in the following table:

Original	Transformed	Original	Transformed
0.32	0.68	1.51	1.15
0.47	0.78	1.62	1.17
0.52	0.80	1.74	1.20
0.59	0.84	1.87	1.23
0.77	0.92	1.89	1.24
0.81	0.93	1.95	1.25
0.81	0.93	2.05	1.27
0.90	0.97	2.10	1.28
0.96	0.99	2.20	1.30
1.18	1.06	2.48	1.35
1.20	1.06	2.81	1.41
1.20	1.06	3.00	1.44
1.31	1.09	3.09	1.46
1.35	1.11	3.37	1.50
1.43	1.13	4.75	1.68

Copyright 2025 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s). Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

Construct a histogram of the transformed data. Compare your histogram to those given in Figure 7.44. Which of the cube-root and square-root transformations appear to result in a histogram that is more nearly symmetric?

7.106 ● The accompanying data on x = student-teacher ratio is for a random sample of 20 high schools in Maine selected from a population of 85 high schools. The data are consistent with summary values for the state of Maine that appeared in an article in the ***Bangor Daily News*** **(bangordailynews.com, September 22, 2016)**. The corresponding normal scores are also shown.

Student-Teacher Ratio (x)	Normal Score
9.0	−1.868
10.0	−1.403
11.0	−1.128
11.2	−0.919
11.6	−0.744
11.7	−0.589
11.8	−0.448
11.9	−0.315
12.0	−0.187
12.1	−0.062
12.5	0.062
12.6	0.187
13.0	0.315
13.2	0.448
13.6	0.589
13.7	0.744
14.0	0.919
14.5	1.128
14.9	1.403
15.0	1.868

a. Construct a normal probability plot.
b. Calculate the correlation coefficient for the (normal score, x) pairs. Based on the value of the correlation coefficient, it is reasonable to think that the distribution of student-teacher ratios for high schools in Maine is approximately normal.

7.107 ● The paper **"Temperature and the Northern Distributions of Wintering Birds" (*Ecology* [1991]: 2274–2285)** gave the following body masses (in grams) for 50 different bird species:

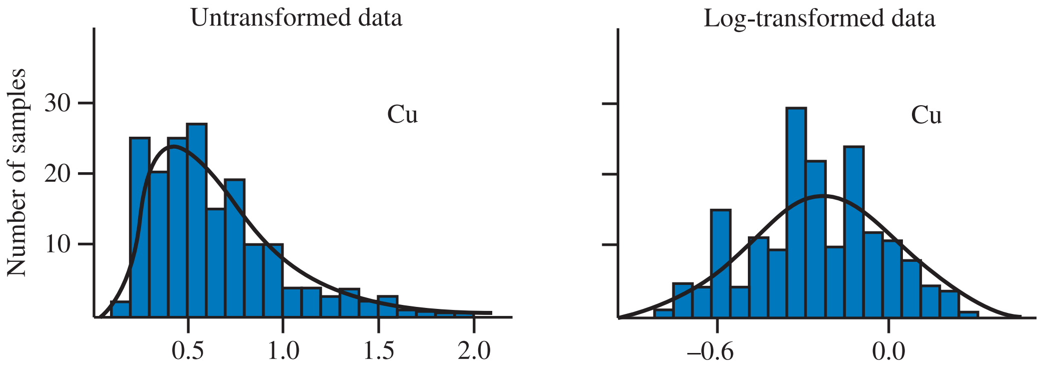
7.7	10.1	21.6	8.6	12.0	11.4	16.6	9.4
11.5	9.0	8.2	20.2	48.5	21.6	26.1	6.2
19.1	21.0	28.1	10.6	31.6	6.7	5.0	68.8
23.9	19.8	20.1	6.0	99.6	19.8	16.5	9.0
448.0	21.3	17.4	36.9	34.0	41.0	15.9	12.5
10.2	31.0	21.5	11.9	32.5	9.8	93.9	10.9
19.6	14.5						

a. Draw a histogram based on class intervals 5 to $<$ 10, 10 to $<$ 15, 15 to $<$ 20, 20 to $<$ 25, 25 to $<$ 30, 30 to $<$ 40, 40 to $<$ 50, 50 to $<$ 100, and 100 to $<$ 500. Is a transformation of the data desirable? Explain.
b. Use a calculator or statistical software package to calculate logarithms of these observations, and construct a histogram. Is the log transformation successful in producing a more symmetric distribution?
c. Consider transformed value $= \dfrac{1}{\sqrt{\text{original value}}}$ and construct a histogram of the transformed data. Does the histogram appear to resemble a normal distribution?

7.108 The figure on the next page appeared in the paper **"EDTA-Extractable Copper, Zinc, and Manganese in Soils of the Canterbury Plains" (*New Zealand Journal of Agricultural Research* [1984]: 207–217)**. A large number of topsoil samples were analyzed for manganese (Mn), zinc (Zn), and copper (Cu), and the resulting data were summarized using histograms.

The investigators transformed each data set using logarithms in an effort to obtain more nearly symmetric distributions of values. Do you think the transformations were successful? Explain.

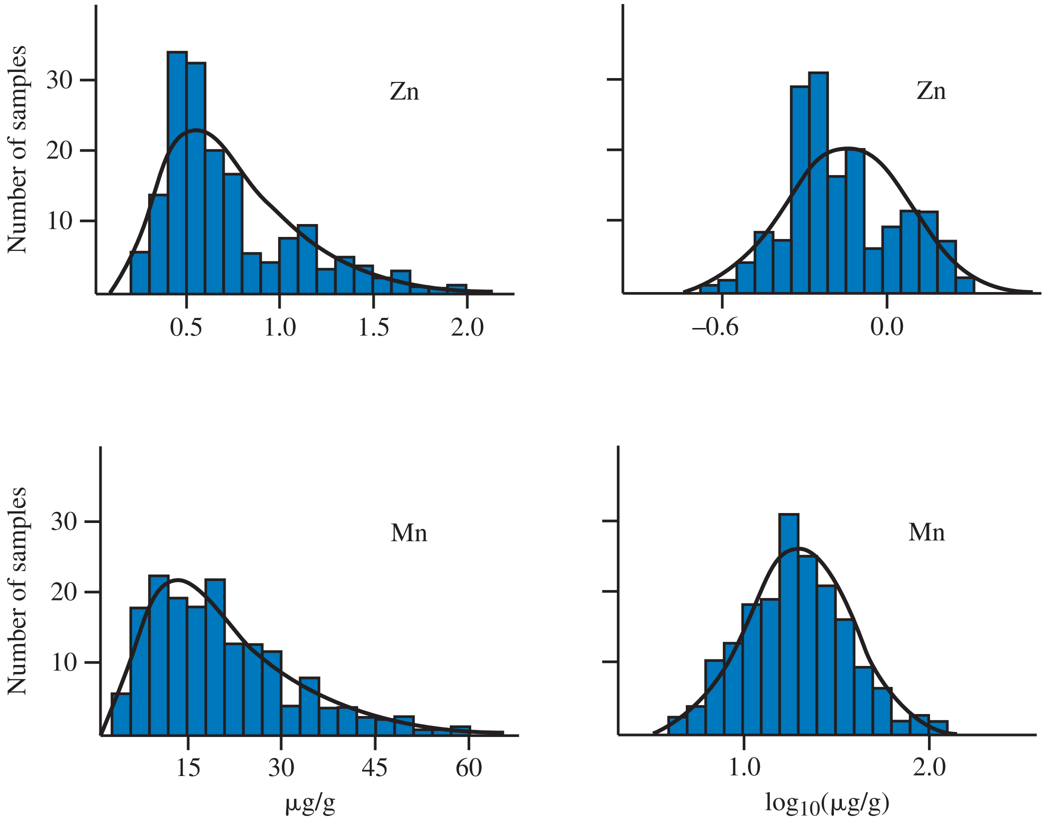
Figure for Exercise 7.108



(continued)

Copyright 2025 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s). Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

Figure for Exercise 7.108 continued



Section 7.8

Using the Normal Distribution to Approximate a Discrete Distribution (Optional)

The distributions of many random variables can be approximated by a carefully chosen normal distribution. In this section, you will see how probabilities for some discrete random variables can be approximated using a normal curve. The most important case of this is the approximation of binomial probabilities.

The Normal Curve and Discrete Variables

The probability distribution of a discrete random variable x is represented graphically by a probability histogram. Possible values of x are isolated points on the number line, usually whole numbers. The probability of a particular value is equal to the area of the rectangle in the histogram that is centered at that value.

Sometimes a probability histogram can be well approximated by a normal curve, as illustrated in Figure 7.49. In such cases, it is customary to say that the distribution of x is approximately normal. The normal distribution can then be used to calculate approximate probabilities of events involving x.

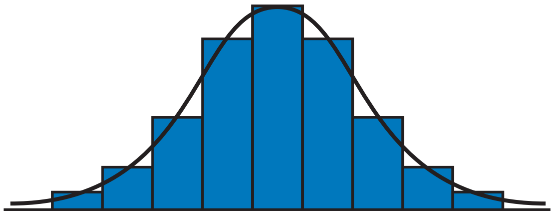


Figure 7.49
A normal curve approximation to a probability histogram.

Copyright 2025 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s). Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

Example 7.37 Express Mail Packages

Understand the context ➤

Suppose the number of express mail packages mailed at a certain post office on a randomly selected day is approximately normally distributed with mean 18 and standard deviation 6. Let's look at how to approximate the probability that $x = 20$.

Figure 7.50(a) shows a portion of the probability histogram for x with the approximating normal curve superimposed. The area of the shaded rectangle is equal to $P(x = 20)$. The left edge of this rectangle is at 19.5 on the horizontal scale, and the right edge is at 20.5. Therefore, the desired probability is approximately the area under the normal curve between 19.5 and 20.5.

Standardizing these limits gives

$$\frac{20.5 - 18}{6} = 0.42 \qquad \frac{19.5 - 18}{6} = 0.25$$

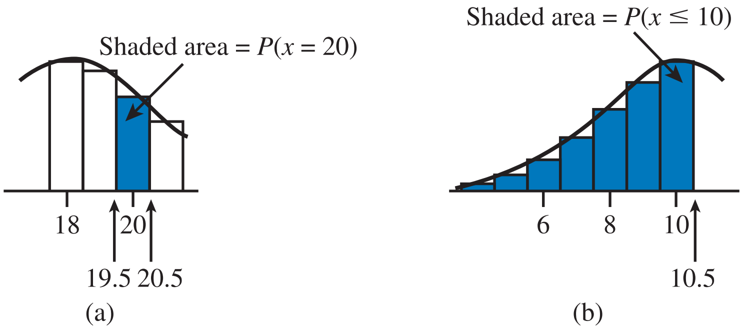
so

$$P(x = 20) \approx P(0.25 < z < 0.42) = 0.6628 - 0.5987 = 0.0641$$

In a similar fashion, Figure 7.50(b) shows that $P(x \le 10)$ is approximately the area under the normal curve to the left of 10.5. Then

$$\begin{aligned} P(x \le 10) \approx P\left(z \le \frac{10.5 - 18}{6}\right) &= P(z \le -1.25) \\ &= 0.1056 \end{aligned}$$

Figure 7.50
The normal approximation for Example 7.37



(a)

(b)

The calculation of probabilities in Example 7.37 illustrates the use of what is known as a **continuity correction**. Because the rectangle for $x = 10$ extends to 10.5 on the right, you use the normal curve area to the left of 10.5 rather than 10. In general, if possible x values are consecutive whole numbers, then $P(a \le x \le b)$ will be approximately the normal curve area between limits $a - \frac{1}{2}$ and $b + \frac{1}{2}$.

Normal Approximation to a Binomial Distribution

Figure 7.51 shows the probability histograms for two binomial distributions, one with $n = 25$, $p = 0.4$, and the other with $n = 25$, $p = 0.1$. For each distribution, $\mu = np$ and $\sigma = \sqrt{np(1-p)}$ were calculated and then a normal curve with the corresponding μ and σ was drawn over the probability histogram.

A normal curve fits the probability histogram well in the first case (Figure 7.51(a)). When this happens, it is reasonable to approximate binomial probabilities using areas under the normal curve.

In the second case (Figure 7.51(b)), the normal curve does not give a good approximation because the probability histogram is skewed, whereas the normal curve is symmetric.

Copyright 2025 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s). Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

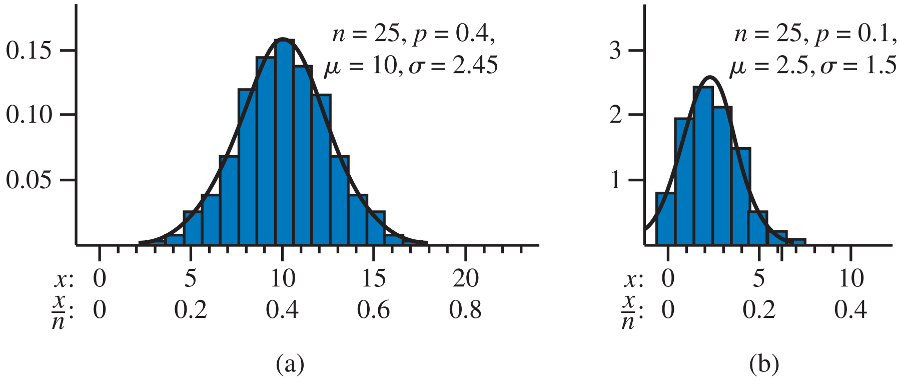


Figure 7.51
Normal approximations to binomial distributions.

Suppose x is a binomial random variable based on n trials and success probability p, so that

$$\mu = np \text{ and } \sigma = \sqrt{np(1-p)}$$

If n and p are such that

$$np \geq 10 \text{ and } n(1-p) \geq 10$$

then the distribution of x is approximately normal.

Combining this result with the continuity correction implies that

$$P(a \leq x \leq b) \approx P\left(\frac{a - \frac{1}{2} - \mu}{\sigma} \leq z \leq \frac{b + \frac{1}{2} - \mu}{\sigma}\right)$$

That is, the probability that x is between a and b inclusive is approximately the area under the approximating normal curve between $a - \frac{1}{2}$ and $b + \frac{1}{2}$.

Similarly,

$$P(x \leq b) \approx P\left(z \leq \frac{b + \frac{1}{2} - \mu}{\sigma}\right) \qquad P(x \geq a) \approx P\left(\frac{a - \frac{1}{2} - \mu}{\sigma} \leq z\right)$$

When either $np < 10$ or $n(1-p) < 10$, the binomial distribution is too skewed for the normal approximation to provide reasonable approximations to binomial probabilities.

Example 7.38 Premature Babies

Understand the context ➤

Premature babies are those born before 37 weeks. According to the Centers for Disease Control and Prevention, 10.1% of babies born in the United States are born before 37 weeks. Suppose that 1000 births will be randomly selected and that the value of

$$x = \text{number of these births that occurred prior to 37 weeks}$$

will be determined. Then x has a binomial distribution with $n = 1000$ and $p = 0.101$. Because

$$np = 1000(0.101) = 101 \geq 10$$
$$n(1-p) = 1000(0.899) = 899 \geq 10$$

the distribution of x is approximately normal with

$$\mu = np = 1000(0.101) = 101$$

Figure 7.52 shows the binomial distribution of x and the approximating normal curse is also shown.

Copyright 2025 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s). Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

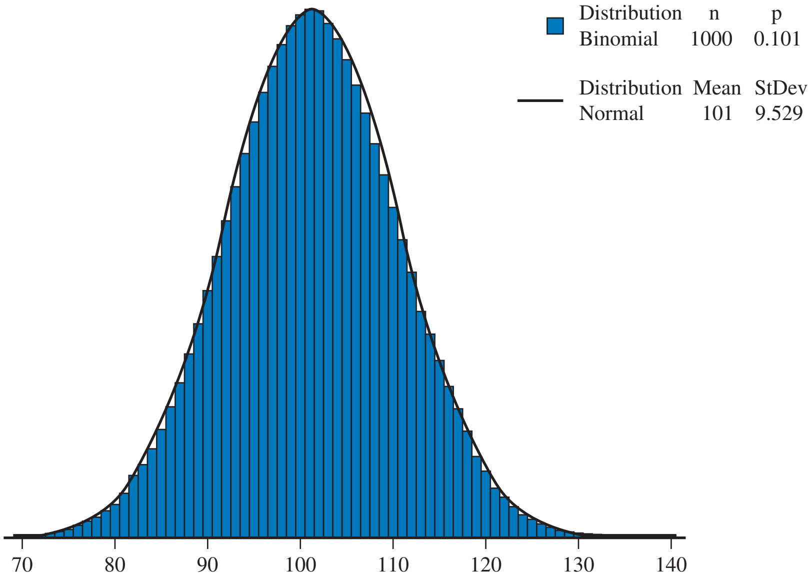


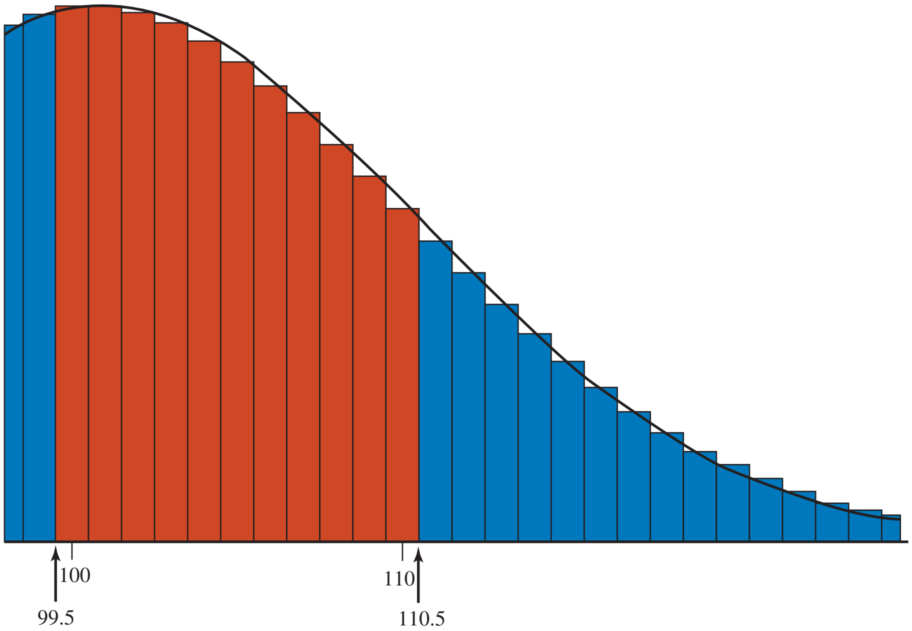
Figure 7.52
Binomial distribution and approximating normal curve for the number of premature births in a sample of size 1000.

Do the work ➤

You can use the normal approximation to calculate probabilities. For example, the probability that a sample of 1000 births will include between 100 and 110 (inclusive) births that occurred before 37 weeks is

$$P(100 \le x \le 110) \approx \text{area under the normal curve between } 99.5 \text{ and } 110.5 = 0.40315$$

as shown in the following figure:



Exercises 7.109–7.120

7.109 Let x denote the score of an individual selected at random from all those taking a national licensing exam. The value of x must be a whole number. Suppose that the distribution of x can be approximated by a normal distribution with mean value 100 and standard deviation 15. Approximate the following probabilities. (Hint: See Example 7.37.)

a. $P(x = 100)$
b. $P(x \le 110)$
c. $P(x < 110)$ (Hint: $x < 110$ is the same as $x \le 109$.)
d. $P(75 \le x \le 125)$

Copyright 2025 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s). Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

7.110 Suppose that the distribution of

$x =$ the number of items produced by an assembly line during an 8-hour shift

can be approximated by a normal distribution with mean value 150 and standard deviation 10.

a. What is the approximate probability that the number of items produced is at most 120?

b. What is the approximate probability that at least 125 items are produced?

c. What is the approximate probability that between 135 and 160 (inclusive) items are produced?

7.111 Suppose that a particular candidate in a local election is favored by a slight majority of voters, with 52% of the eligible voters favoring this candidate. If a random sample of 100 eligible voters is selected, what is the approximate probability that fewer than half the voters in the sample favor this candidate?

7.112 Suppose that x has a binomial distribution with $n = 50$ and $p = 0.6$, so that $\mu = np = 30$ and $\sigma = \sqrt{np(1 - p)} = 3.464$. Approximate the following probabilities using the normal approximation with the continuity correction. (Hint: See Example 7.37.)

a. $P(x = 30)$

b. $P(x = 25)$

c. $P(x \le 25)$

7.113 For the binomial distribution described in the previous exercise, approximate the following probabilities using the normal approximation with the continuity correction.

a. $P(25 \le x \le 40)$

b. $P(25 < x < 40)$ (Hint: $25 < x < 40$ is the same as $26 \le x \le 39$.)

7.114 Symptom validity tests (SVTs) are sometimes used to confirm diagnosis of psychiatric disorders. The paper **"Developing a Symptom Validity Test for Post-Traumatic Stress Disorder: Application of the Binomial Distribution" (*Journal of Anxiety Disorders* [2008]: 1297–1302)** investigated the use of SVTs in the diagnosis of post-traumatic stress disorder.

One SVT proposed is a 60-item test (called the MENT test), where each item has only a correct or incorrect response. The MENT test is designed so that responses to the individual questions can be considered independent of one another. For this reason, the authors of the paper believe that the score on the MENT test can be viewed as a binomial random variable with $n = 60$.

The MENT test is designed to help in distinguishing fake claims of post-traumatic stress disorder. The items on the MENT test are written so that the correct response to an item should be relatively obvious, even to people suffering from stress disorders. Researchers have found that a patient with a fake claim of stress disorder will try to fake the test, and that the probability of a correct response to an item for these patients is 0.7 (compared to 0.96 for other patients).

The authors used a normal approximation to the binomial distribution with $n = 60$ and $p = 0.7$ to approximate various probabilities of interest, where

$x =$ number of correct responses on the MENT test for a patient who is trying to fake the test.

a. Verify that it is appropriate to use a normal approximation to the binomial distribution in this situation.

b. Approximate the following probabilities:

- **i.** $P(x = 42)$
- **ii.** $P(x < 42)$
- **iii.** $P(x \le 42)$

c. Explain why the probabilities computed in Part (b) are not all equal.

7.115 Consider the information on the MENT test given in the previous exercise.

a. The exact binomial probability of a score of 42 or less for someone who is not faking the test ($p = 0.96$) is

$p(x \le 42) = 0.000000000013$

Explain why this probability was calculated using the binomial formula rather than using a normal approximation.

b. The authors of the study described in the previous exercise propose that someone who scores 42 or less on the MENT exam is faking the test. Explain why this is reasonable, using the probability from Part (a) and the probabilities from the previous exercise as justification.

7.116 Suppose that 70% of the bicycles sold by a certain store are mountain bikes. Among 100 randomly selected bike purchases, what is the approximate probability that

a. At most 75 are mountain bikes?

b. Between 60 and 75 (inclusive) are mountain bikes?

c. More than 80 are mountain bikes?

d. At most 30 are not mountain bikes?

7.117 Suppose that 25% of the fire alarms in a large city are false alarms. Let x denote the number of false alarms in a random sample of 100 alarms. Approximate the following probabilities:

a. $P(20 \le x \le 30)$

b. $P(20 < x < 30)$

c. $P(x \ge 35)$

d. The probability that x is farther than 2 standard deviations from its mean value

7.118 Suppose that 65% of all registered voters in a certain area favor a 7-day waiting period before purchase of a handgun. Among 225 randomly

Copyright 2025 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s). Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

selected registered voters, what is the approximate probability that

- **a.** At least 150 favor such a waiting period?
- **b.** More than 150 favor such a waiting period?
- **c.** Fewer than 125 favor such a waiting period?

7.119 Flashlight bulbs manufactured by a certain company are sometimes defective.

- **a.** If 5% of all such bulbs are defective, could the techniques of this section be used to approximate the probability that at least five of the bulbs in a random sample of size 50 are defective? If so, calculate this probability; if not, explain why not.
- **b.** Reconsider the question posed in Part (a) for the probability that at least 20 bulbs in a random sample of size 500 are defective.

7.120 A company that manufactures mufflers for cars offers a lifetime warranty on its products, provided that ownership of the car does not change. Suppose that only 20% of its mufflers are replaced under this warranty.

- **a.** In a random sample of 400 purchases, what is the approximate probability that between 75 and 100 (inclusive) mufflers are replaced under warranty?
- **b.** Among 400 randomly selected purchases, what is the approximate probability that at most 70 mufflers are ultimately replaced under warranty?
- **c.** If you were told that fewer than 50 among 400 randomly selected purchases were ever replaced under warranty, would you question the 20% figure? Explain.

Chapter Activities

Activity 7.1 Is It Real?

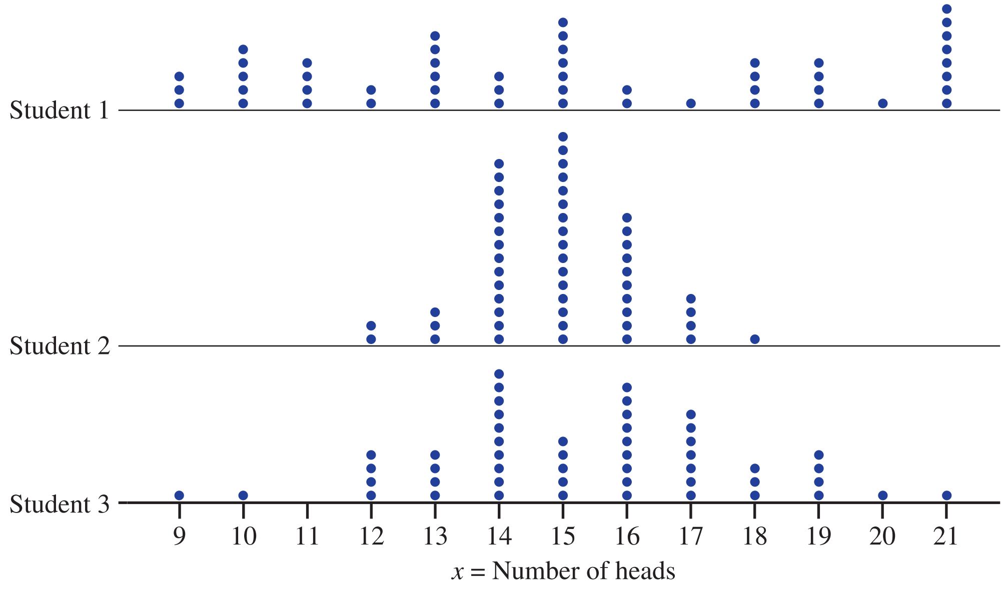
The purpose of this activity is to provide you with an opportunity to confirm your understanding of how probability distributions describe long-run behavior of a random variable.

Background: Three students were asked to complete an assignment that requested they do the following:

- **a.** Flip a coin 30 times, and note the number of heads observed in the 30 flips.
- **b.** Repeat Step (a) 100 times, to obtain 100 observations of the random variable x = number of heads in 30 flips.
- **c.** Construct a dotplot of the 100 x values.

Because this was a tedious assignment, one or more of the students did not really carry out the coin flipping and just made up 100 values of x that they thought would look "real." The dotplots produced by these three students are shown below.

1. Do you think that any of the three students made up the x values shown in their dotplot? If so, which ones, and what about the dotplot makes you think the student did not actually do the coin flipping?
2. Working as a group, each student in your class should flip a coin 30 times and note the number of heads in the 30 tosses. If there are fewer than 50 students in the class, each student should repeat this process until there are a total of at least 50 observations of x = number of heads in 30 flips. Using the data from the entire class, construct a dotplot of the x values.
3. After looking at the dotplot from Step 2 that resulted from actually flipping a coin 30 times and observing number of heads, reconsider your answers in Question 1. For each of the three students, explain why you now think that they did or did not actually do the coin flipping.



Copyright 2025 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s). Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

Activity 7.2 Rotten Eggs?

The purpose of this activity is to use what you have learned about probability distributions and simulation to evaluate a claim.

Background: *The Salt Lake Tribune* (October 11, 2002) printed the following account of an exchange between a restaurant manager and a health inspector:

> The recipe calls for four fresh eggs for each quiche. A Salt Lake County Health Department inspector paid a visit recently and pointed out that research by the Food and Drug Administration indicates that one in four eggs carries salmonella bacterium, so restaurants should never use more than three eggs when preparing quiche. The manager on duty wondered aloud if simply throwing out three eggs from each dozen and using the remaining nine in four-egg quiches would serve the same purpose.

1. Working in a group or as a class, discuss the folly of the above statement!

2. Suppose the following argument is made for three-egg quiches rather than four-egg quiches: Let $x =$ number of eggs that carry salmonella. Then

   $$p(0) = p(x = 0) = (0.75)^3 = 0.422$$

   for three-egg quiches and

   $$p(0) = p(x = 0) = (0.75)^4 = 0.316$$

   for four-egg quiches. What assumption must be made to justify these probability calculations? Do you think this is reasonable or not? Explain.

3. Suppose that a carton of one dozen eggs does happen to have exactly three eggs that carry salmonella and that the manager does as he proposes: selects three eggs at random and throws them out, then uses the remaining nine eggs in four-egg quiches. Let $x =$ number of eggs that carry salmonella among four eggs selected at random from the remaining nine.

   Working with a partner, conduct a simulation to approximate the distribution of x by carrying out the following sequence of steps:

   a. Take 12 identical slips of paper and write “Good” on nine of them and “Bad” on the remaining three. Place the slips of paper in a paper bag or some other container.
   b. Mix the slips and then select three at random and remove them from the bag.
   c. Mix the remaining slips and select four “eggs” from the bags.
   d. Note the number of bad eggs among the four selected. (This is an observed x value.)
   e. Replace all slips, so that the bag now contains all 12 “eggs.”
   f. Repeat Steps (b)–(d) at least 10 times, each time recording the observed x value.

4. Combine the observations from your group with those from the other groups. Use the resulting data to approximate the distribution of x. Comment on the resulting distribution in the context of the risk of salmonella exposure if the manager’s proposed procedure is used.

Summary Key Concepts and Formulas

Term or Formula	Comment
Random variable: discrete or continuous	A numerical variable with a value determined by the outcome of a chance experiment. A random variable is discrete if its possible values are isolated points along the number line and continuous if its possible values form an entire interval on the number line.
Probability distribution of a discrete random variable x	A formula, table, or graph that gives the probability associated with each possible x value. Conditions on $p(x)$ are (1) $p(x) \geq 0$, and (2) $\Sigma p(x) = 1$, where the sum is over all possible x values.
Probability distribution of a continuous random variable x	Specified by a smooth (density) curve for which the total area under the curve is 1. The probability $P(a < x < b)$ is the area under the curve and above the interval from a to b; this is also $P(a \leq x \leq b)$.

Term or Formula	Comment
μ_x **and** σ_x	The mean and standard deviation, respectively, of a random variable x. These quantities describe the center and extent of spread about the center of the variable’s probability distribution.
$\mu_x = \Sigma x p(x)$	The mean value of a discrete random variable x. It locates the center of the variable’s probability distribution.
$\sigma_x^2 = \Sigma(x - \mu_x)^2 p(x)$ $\sigma_x = \sqrt{\sigma_x^2}$	The variance and standard deviation, respectively, of a discrete random variable. These are measures of the extent to which the variable’s distribution spreads out about the mean μ_x.
Binomial probability distribution $p(x) = \dfrac{n!}{x!(n-x)!} p^x (1-p)^{n-x}$	This formula gives the probability of observing x successes ($x = 0, 1, \ldots, n$) among n trials of a binomial experiment.

Copyright 2025 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s). Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

Term or Formula	Comment
$\mu_x = np$ $\sigma_x = \sqrt{np(1-p)}$	The mean and standard deviation of a binomial random variable.
Normal distribution	A continuous probability distribution that has a bell-shaped density curve. A particular normal distribution is determined by specifying values of μ and σ.
Standard normal distribution	This is the normal distribution with $\mu = 0$ and $\sigma = 1$. The density curve is called the z curve, and z is the letter commonly used to denote a variable having this distribution. Areas under the z curve to the left of various values are given in Appendix Table 2.
$z = \frac{x-\mu}{\sigma}$	z is obtained by "standardizing": subtracting the mean and then dividing by the standard deviation. When x has a normal distribution, z has a standard normal distribution.
Normal probability plot	A graph used to judge the plausibility of the assumption that a sample has been selected from a normal population distribution. If the plot is reasonably straight, this assumption is reasonable.
Normal approximation to the binomial distribution	When both $np \geq 10$ and $n(1-p) \geq 10$, binomial probabilities are well approximated by corresponding areas under a normal curve with $\mu = np$ and $\sigma = \sqrt{np(1-p)}$.

Chapter Review Exercises 7.121–7.133

7.121 Let x denote the duration of a randomly selected pregnancy (the time elapsed between conception and birth). Accepted values for the mean value and standard deviation of x are 266 days and 16 days, respectively. Suppose that the probability distribution of x is (approximately) normal.

a. What is the probability that the duration of pregnancy is between 250 and 300 days?

b. What is the probability that the duration of pregnancy is at most 240 days?

c. What is the probability that the duration of pregnancy is within 16 days of the mean duration?

d. Some insurance companies will pay the medical expenses associated with childbirth only if the insurance has been in effect for more than 9 months (275 days). This restriction is designed to ensure that the insurance company pays benefits for only those pregnancies for which conception occurred during coverage. Suppose that conception occurred 2 weeks after coverage began. What is the probability that the insurance company will refuse to pay benefits because of the 275-day insurance requirement?

7.122 A soft-drink machine dispenses only regular Coke and Diet Coke. Sixty percent of all purchases from this machine are diet drinks. The machine currently has 10 cans of each type. If 15 customers want to purchase drinks before the machine is restocked, what is the probability that each of the 15 is able to purchase the type of drink desired? (Hint: Let x denote the number among the 15 who want a diet drink. For which possible values of x is everyone satisfied?)

7.123 A business has six customer service telephone lines. Let x denote the number of lines in use at a specified time. Suppose that the probability distribution of x is as follows:

x	0	1	2	3	4	5	6
$p(x)$	0.10	0.15	0.20	0.25	0.20	0.06	0.04

Write each of the following events in terms of x, and then calculate the probability of each one:

a. At most three lines are in use

b. Fewer than three lines are in use

c. At least three lines are in use

7.124 Use the probability distribution for x = number of lines in use given in the previous exercise to calculate the probability of each of the following events.

a. Between two and five lines (inclusive) are in use

b. Between two and four lines (inclusive) are not in use

c. At least four lines are not in use

7.125 Refer to the probability distribution given in Exercise 7.123.

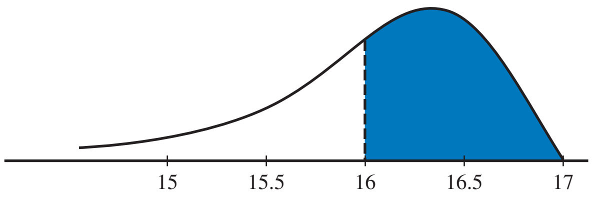
a. Calculate the mean value and standard deviation of x.

b. What is the probability that the number of lines in use is more than 3 standard deviations from the mean value?

7.126 Suppose that the random variable

x = actual weight (in ounces) of a randomly selected package of cereal

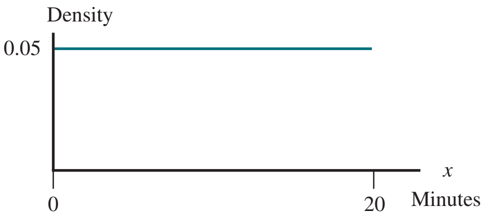
has the probability distribution described by the density curve pictured here.



a. What probability is represented by the shaded area?

b. Suppose the shaded area = 0.50. Interpret this probability in the context of this problem.

7.127 Let x be the amount of time (in minutes) that a particular San Francisco commuter will have to wait for a BART train. Suppose that the probability distribution of x is as pictured (a uniform distribution):



a. What is the probability that x is less than 10 minutes? More than 15 minutes?

b. What is the probability that x is between 7 and 12 minutes?

c. Find the value c for which $P(x < c) = 0.9$.

7.128 Suppose that fuel efficiency for a particular model car under specified conditions is normally distributed with a mean value of 30.0 mpg and a standard deviation of 1.2 mpg.

a. What is the probability that the fuel efficiency for a randomly selected car of this type is between 29 and 31 mpg?

b. Would it surprise you to find that the efficiency of a randomly selected car of this model is less than 25 mpg?

c. If three cars of this model are randomly selected, what is the probability that all three have efficiencies exceeding 32 mpg?

d. Find a number c^* such that 95% of all cars of this model have efficiencies exceeding c^* (that is, $P(x > c^*) = 0.95$).

7.129 A coin is flipped 25 times. Let x be the number of flips that result in heads (H). Consider the following rule for deciding whether or not the coin is fair:

Judge the coin fair if $8 \leq x \leq 17$.
Judge the coin biased if either $x \leq 7$ or $x \geq 18$.

a. What is the probability of judging the coin biased when it is actually fair?

b. What is the probability of judging the coin fair when $P(\text{H}) = 0.9$, so that there is a substantial bias? Repeat for $P(\text{H}) = 0.1$.

c. What is the probability of judging the coin fair when $P(\text{H}) = 0.6$? when $P(\text{H}) = 0.4$? Why are these probabilities so large compared to the probabilities in Part (b)?

d. What happens to the "error probabilities" of Parts (a) and (b) if the decision rule is changed so that the coin is judged fair if $7 \leq x \leq 18$ and unfair otherwise? Is this a better rule than the one first proposed? Explain.

7.130 The probability distribution of x, the number of defective tires on a randomly selected automobile checked at a certain inspection station, is given in the following table:

x	0	1	2	3	4
$p(x)$	0.54	0.16	0.06	0.04	0.20

The mean value of x is $\mu_x = 1.2$. Calculate the values of σ_x^2 and σ_x.

7.131 The amount of time spent by a statistical consultant with a client at their first meeting is a random variable having a normal distribution with a mean value of 60 minutes and a standard deviation of 10 minutes.

a. What is the probability that more than 45 minutes is spent at the first meeting?

b. What amount of time is exceeded by only 10% of all clients at a first meeting?

c. If the consultant assesses a fixed charge of \$10 (for overhead) and then charges \$50 per hour, what is the mean revenue from a client's first meeting?

Copyright 2025 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s). Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

7.132 The lifetime of a certain brand of battery is normally distributed with a mean value of 6 hours and a standard deviation of 0.8 hours when it is used in a particular DVD player. Suppose that two new batteries are independently selected and put into the player. The player stops functioning as soon as one of the batteries fails.

a. What is the probability that the player functions for at least 4 hours?

b. What is the probability that the DVD player works for at most 7 hours?

c. Find a number c^* such that only 5% of all DVD players will function without battery replacement for more than c^* hours.

7.133 The longest "run" of S's in the sequence $SSFSSSSFFS$ has length 4, corresponding to the S's in the fourth, fifth, sixth, and seventh positions. Consider a binomial experiment with $n = 4$, and let y be the length in the longest run of S's.

a. When $p = 0.5$, the 16 possible outcomes are equally likely. Determine the probability distribution of y in this case (first list all outcomes and the y value for each one). Then calculate μ_y.

b. Repeat Part (a) for the case $p = 0.6$.

c. Let z denote the longest run of either S's or F's. Determine the probability distribution of z when $p = 0.5$.

7.134 Four people—a, b, c, and d—are waiting to give blood. Of these four, a and b have type AB blood, whereas c and d do not. An emergency call has just come in for some type AB blood. If blood donations are taken one by one from the four people in random order and x is the number of donations needed to obtain an AB individual (so possible x values are 1, 2, and 3), what is the probability distribution of x?

7.135 Kyle and Lygia are going to play a series of chess games. The first person to win four games will be declared the winner. Suppose that outcomes of successive games are independent and that the probability of Lygia winning any particular game is 0.6. Define a random variable x as the number of games played in the series.

a. What is $p(4)$? (Hint: Either Kyle or Lygia could win four straight games.)

b. What is $p(5)$? (Hint: For Lygia to win in exactly five games, what has to happen in the first four games and in Game 5?)

c. Determine the probability distribution of x.

d. On average, how many games will the series last?

7.136 Suppose that your statistics professor tells you that the scores on a midterm exam were approximately normally distributed with a mean of 78 and a standard deviation of 7. The top 15% of all scores have been designated A's. Your score is 89. Did you receive an A? Explain.

7.137 Suppose that the pH of soil samples taken from a certain geographic region is normally distributed with a mean pH of 6.00 and a standard deviation of 0.10. If the pH of a randomly selected soil sample from this region is determined, answer the following questions about it:

a. What is the probability that the resulting pH is between 5.90 and 6.15?

b. What is the probability that the resulting pH exceeds 6.10?

c. What is the probability that the resulting pH is at most 5.95?

d. What value will be exceeded by only 5% of all such pH values?

7.138 The lightbulbs used to provide exterior lighting for a large office building have an average lifetime of 700 hours. If length of life is approximately normally distributed with a standard deviation of 50 hours, how often should all the bulbs be replaced so that no more than 20% of the bulbs will have already burned out?

7.139 Suppose there are approximately 40,000 travel agencies in the United States, of which 11,000 are members of the American Society of Travel Agents (ASTA).

a. If x is the number of ASTA members among 5000 randomly selected agencies, could you use the methods of Section 7.8 to approximate $P(1200 < x < 1400)$? Why or why not?

b. In a random sample of 100 agencies, what are the mean value and standard deviation of the number of ASTA members?

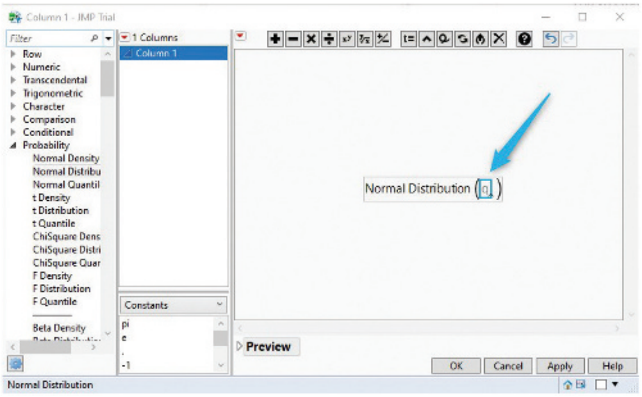
c. If the sample size in Part (b) is doubled, does the standard deviation double? Explain.

Technology Notes

Finding Normal Probabilities

JMP

1. After opening a new data table, click **Rows** then select **Add rows**
2. Type **1** in the box next to **How many rows to add:**
3. Click **OK**
4. Double-click on the **Column 1** heading
5. Click **Column Properties** and select **Formula**
6. Click **Edit Formula**
7. In the left panel, expand the **Probability** menu then select **Normal Distribution**
8. In the right panel, double-click in the blue box around *q*


Column 1 - JMP Trial
Filter
Row
Numeric
Transcendental
Trigonometric
Character
Comparison
Conditional
Probability
Normal Density
Normal Distribu
Normal Quantil
t Density
t Distribution
t Quantile
ChiSquare Dens
ChiSquare Distri
ChiSquare Quar
F Density
F Distribution
F Quantile
Beta Density
1 Columns
Column 1
Constants
pi
e
.
-1
Normal Distribution (q)
Preview
OK
Cancel
Apply
Help
Normal Distribution

Source: JMP

9. Type the value that you would like to find a probability from the standard normal distribution and click **OK**
10. Click **OK**

Note: This procedure outputs the value for $P(X \leq x)$. If you want to find the value for $P(X \geq x)$ you will need to subtract the output from one.

Note: You may also type a column of values for which you would like to find probabilities (these must ALL be from the SAME distribution). On Step 9, instead of typing one value, click and drag the column name into the blue box.

Minitab

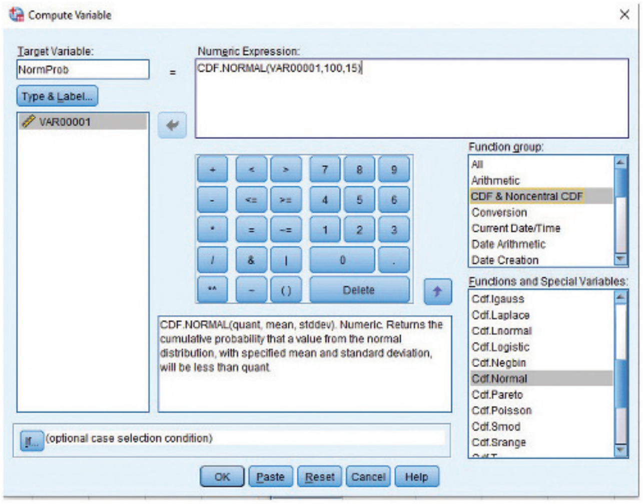
1. Click **Calc** then click **Probability Distributions** then click **Normal…**
2. In the box next to **Mean:** type the mean of the normal distribution that you are working with
3. In the box next to **Standard deviation:** type the standard deviation of the normal distribution that you are working with
4. Click the radio button next to **Input constant:**
5. Click in the box next to **Input constant:** and type the value that you are finding the probability for
6. Click **OK**

Note: This procedure outputs the value for $P(X \leq x)$. If you want to find the value for $P(X \geq x)$ you will need to subtract the output from one.

Note: You may also type a column of values for which you would like to find probabilities (these must ALL be from the SAME distribution) and use the **Input Column** option to find probabilities for each value in the column selected.

SPSS

1. Type in the values that you would like to find the probabilities for in one column (this column will be automatically titled VAR00001)
2. Click **Transform** and select **Compute Variable…**
3. In the box under **Target Variable:** type NormProb (this will be the title of a new column that lists the cumulative probabilities for each value in VAR00001)
4. Under **Function group:** select **CDF & Noncentral CDF**
5. Under **Functions and Special Variables:** double-click **Cdf. Normal**
6. In the box under **Numeric Expression:** you should see CDF. NORMAL(?,?,?)
7. Highlight the first ? and double-click VAR00001
8. Highlight the second ? and type the mean of the normal distribution that you are working with
9. Highlight the third ? and type the standard deviation of the normal distribution that you are working with


Target Variable: NormProb

Numeric Expression: CDF.NORMAL(VAR00001,100,15)

CDF.NORMAL(quant, mean, stddev). Numeric. Returns the cumulative probability that a value from the normal distribution, with specified mean and standard deviation, will be less than quant.

Function group: CDF & Noncentral CDF

Functions and Special Variables: Cdf.Normal

Compute Variable
Type & Label...
VAR00001
=
+ < > 7 8 9
- <= >= 4 5 6
* = ~= 1 2 3
/ & | 0 .
** ~ ( ) Delete
All
Arithmetic
CDF & Noncentral CDF
Conversion
Current Date/Time
Date Arithmetic
Date Creation
Cdf.Igauss
Cdf.Laplace
Cdf.Lnormal
Cdf.Logistic
Cdf.Negbin
Cdf.Normal
Cdf.Pareto
Cdf.Poisson
Cdf.Smod
Cdf.Srange
If...
(optional case selection condition)
OK
Paste
Reset
Cancel
Help

Source: SPSS

10. Click **OK**

Note: This procedure outputs the value for $P(X \leq x)$. If you want to find the value for $P(X \geq x)$ you will need to subtract the output from one.

Excel

1. Click in an empty cell
2. Click the **Formulas** ribbon
3. Select **Insert Function**
4. Select **Statistical** from the drop-down menu for category
5. Select **NORMDIST** from the **Select a function:** box
6. Click **OK**

7. Click in the box next to **X** and type the data value that you would like to find the probability for
8. Click in the box next to **Mean** and type the mean of the Normal distribution that you are working with
9. Click in the box next to **Standard_dev** and type the standard deviation of the Normal distribution that you are working with
10. Click in the box next to **Cumulative** and type **TRUE**
11. Click **OK**

Note: This procedure outputs the value for $P(X \leq x)$. If you want to find the value for $P(X \geq x)$ you will need to subtract the output from one.

TI-83/84

$P(x_1 < X < x_2)$

1. Press the **2nd** key then press the **VARS** key
2. Highlight **normalcdf(** and press **ENTER**
3. Type in the lower bound x_1
4. Type a comma if using a TI-83; press **ENTER** is using a TI-84
5. Type in the upper bound x_2
6. Type a comma if using a TI-83; press **ENTER** is using a TI-84
7. Type in the value of the mean
8. Type a comma if using a TI-83; press **ENTER** is using a TI-84
9. Type in the value of the standard deviation
10. Type a right parenthesis if using a TI-83; press **ENTER** is using a TI-84
11. Press **ENTER** once if using a TI-83; Press **ENTER** twice if using a TI-84

Note: If you are looking for $P(X < x)$, use a lower bound of $-10{,}000{,}000$. If you are looking for $P(X > x)$, use an upper bound of 10,000,000.

TI-Nspire

$P(x_1 < X < x_2)$

1. Under **Scratchpad** select **A Calculate**
2. Press the **menu** key and select **5:Probablity** then **5:Distributions** then **2:Normal Cdf…** then press **enter**
3. In the box next to **Lower Bound** type the value for x_1
4. In the box next to **Upper Bound** type the value for x_2
5. In the box next to μ type the value of the mean
6. In the box next to σ type the value of the standard deviation
7. Press **OK**

Note: If you are looking for $P(X < x)$, use a lower bound of $-10{,}000{,}000$. If you are looking for $P(X > x)$, use an upper bound of 10,000,000.

Finding Binomial Probabilities

JMP

$P(X \leq x)$

1. After opening a new data table, click **Rows** then select **Add rows**
2. Type **1** in the box next to **How many rows to add:**
3. Click **OK**
4. Double-click on the **Column 1** heading
5. Click **Column Properties** and select **Formula**
6. Click **Edit Formula**
7. In the left panel, expand the **Discrete Probability** menu then select **Binomial Distribution**
8. In the right panel, double-click in the box around p and type the value for the success probability, p, and press **enter** on your keyboard
9. Double-click in the box around n and type the value for the number of trials, n, and press **enter** on your keyboard
10. Double-click in the box around k and type the value for the number of successes for which you would like to find the probability and press **enter** on your keyboard
11. Click **OK**
12. Click **OK**

$P(X = x)$

1. After opening a new data table, click **Rows** then select **Add rows**
2. Type **1** in the box next to **How many rows to add:**
3. Click **OK**
4. Double-click on the **Column 1** heading
5. Click **Column Properties** and select **Formula**
6. Click **Edit Formula**
7. In the left panel, expand the **Discrete Probability** menu then select **Binomial Probability**
8. In the right panel, double-click in the box around p and type the value for the success probability, p, and press **enter** on your keyboard
9. Double-click in the box around n and type the value for the number of trials, n, and press **enter** on your keyboard
10. Double-click in the box around k and type the value for the number of successes for which you would like to find the probability and press **enter** on your keyboard
11. Click **OK**
12. Click **OK**

Minitab

$P(X \leq x)$

1. Click **Calc** then click **Probability Distributions** then click **Binomial…**
2. In the box next to **Number of Trials:** input the value for n, the total number of trials
3. In the box next to **Probability of Success:** input the value for p, the success probability
4. Click the radio button next to **Input Constant**
5. In the box next to **Input Constant:** type the value that you want to find the probability for
6. Click **OK**

Note: You may also type a column of values for which you would like to find probabilities (these must ALL be from the SAME distribution) and use the **Input Column** option to find probabilities for each value in the column selected.

$P(X = x)$

1. Click **Calc** then click **Probability Distributions** then click **Binomial…**
2. Click the radio button next to **Probability**
3. In the box next to **Number of Trials:** input the value for n, the total number of trials

Copyright 2025 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s). Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

4. In the box next to **Probability of Success:** input the value for p, the success probability
5. Click the radio button next to **Input Constant**
6. In the box next to **Input Constant:** type the value that you want to find the probability for
7. Click **OK**

Note: You may also type a column of values for which you would like to find probabilities (these must ALL be from the SAME distribution) and use the **Input Column** option to find probabilities for each value in the column selected.

SPSS

$P(X \le x)$

1. Type in the values that you would like to find the probabilities for in one column (this column will be automatically titled VAR00001)
2. Click **Transform** and select **Compute Variable…**
3. In the box under **Target Variable:** type **BinomProb** (this will be the title of a new column that lists the cumulative probabilities for each value in VAR00001)
4. Under **Function group:** select **CDF & Noncentral CDF**
5. Under **Functions and Special Variables:** double-click **Cdf. Binom**
6. In the box under **Numeric Expression:** you should see CDF. BINOM(?,?,?)
7. Highlight the first ? and double-click VAR00001
8. Highlight the second ? and type the value for n, the total number of trials
9. Highlight the third ? and type success probability, p
10. Click **OK**

$P(X = x)$

1. Type in the values that you would like to find the probabilities for in one column (this column will be automatically titled VAR00001)
2. Click **Transform** and select **Compute Variable…**
3. In the box under **Target Variable:** type BinProb (this will be the title of a new column that lists the cumulative probabilities for each value in VAR00001)
4. Under **Function group:** select **PDF & Noncentral PDF**
5. Under **Functions and Special Variables:** double-click Pdf. Binom
6. In the box under **Numeric Expression:** you should see PDF. BINOM(?,?,?)
7. Highlight the first ? and double-click VAR00001
8. Highlight the second ? and type the value for n, the total number of trials
9. Highlight the third ? and type success probability, p
10. Click **OK**

Excel

1. Click in an empty cell
2. Click the **Formulas** ribbon
3. Select **Insert Function**
4. Select **Statistical** from the drop-down menu for category
5. Select **BINOMDIST** from the **Select a function:** box
6. Click **OK**
7. Click in the box next to **Number_s** and type the number of successes that you are finding a probability for
8. Click in the box next to **Trials** and type the value for n, the total number of trials
9. Click in the box next to **Probability_s** and type success probability, p
10. Click in the box next to **Cumulative** and type **TRUE** if you are finding $P(X \le x)$ or type **FALSE** if you are finding $P(X = x)$
11. Click **OK**

Note: This procedure outputs the value for $P(X \le x)$ when **TRUE** is used as input for **Cumulative**. If you want to find the value for $P(X \ge x)$ you will need to subtract this output from one.

TI-83/84

$P(X \le x)$

1. Press the **2nd** key then press the **VARS** key
2. Highlight the **binomcdf(** option and press the **ENTER** key
3. Type in the number of trials, n
4. Type a comma if using a TI-83; press **ENTER** is using a TI-84
5. Type in the success probability, p
6. Type a comma if using a TI-83; press **ENTER** is using a TI-84
7. Type the value for x
8. Type a right parenthesis if using a TI-83; press **ENTER** is using a TI-84
9. Press **ENTER** once if using a TI-83; Press **ENTER** twice if using a TI-84

$P(X = x)$

1. Press the **2nd** key then press the **VARS** key
2. Highlight the **binompdf(** option and press the **ENTER** key
3. Type in the number of trials, n
4. Type a comma if using a TI-83; press **ENTER** is using a TI-84
5. Type in the success probability, p
6. Type a comma if using a TI-83; press **ENTER** is using a TI-84
7. Type the value for x
8. Type a right parenthesis if using a TI-83; press **ENTER** is using a TI-84
9. Press **ENTER** once if using a TI-83; Press **ENTER** twice if using a TI-84

TI-Nspire

$P(x_1 \le X \le x_2)$

1. Under **Scratchpad** select **A Calculate**
2. Press the **menu** key and select **5:Probability** then select **5:Distributions** then select **E:Binomial Cdf…** and press the **enter** key
3. In the box next to **Num Trials, n** type in the number of trials, n
4. In the box next to **Prob Success, p** type in the success probability, p
5. In the box next to **Lower Bound** type in the value for x_1
6. In the box next to **Upper Bound** type in the value for x_2
7. Press **OK**

Note: In order to find $P(X \le x)$ input a 0 for the lower bound.

$P(X = x)$

1. Under **Scratchpad** select **A Calculate**

2. Press the **menu** key and select **5:Probability** then select **5:Distributions** then select **D:Binomial Pdf…** and press the **enter** key
3. In the box next to **Num Trials, n** type in the number of trials, n
4. In the box next to **Prob Success, p** type in the success probability, p
5. In the box next to **X Value** type the value for x
6. Press **OK**

Normal Probability Plots

JMP

1. Enter the raw data into a column
2. Click **Analyze** then select **Distribution**
3. Click and drag the column name containing the data from the box under **Select Columns** to the box next to **Y, Columns**
4. Click **OK**
5. Click the red arrow next to the column name
6. Select **Normal Quantile Plot**

Minitab

1. Input the data for which you would like to check Normality into a column
2. Click **Graph** then click **Probability Plot…**
3. Highlight the **Single** plot
4. Click **OK**
5. Double click the column name for the column that contains your data to move it into the **Graph Variables:** box
6. Click **OK**

SPSS

1. Input the data for which you would like to check Normality into a column
2. Click **Analyze** then select **Descriptive Statistics** then select **Q-Q Plots…**
3. Highlight the column name for the variable
4. Click the arrow to move the variable to the **Variables:** box
5. Click **OK**

Note: The normal probability plot is output with several other plots and statistics. This plot can be found under the title **Normal Q-Q Plot**.

Excel

Excel does not have the functionality to automatically produce Normal Probability Plots.

However, Excel can produce Normal Probability Plots in the course of running a regression using the Analysis ToolPak.

TI-83/84

1. Input the raw data into **L1** (In order to access lists press the **STAT** key, highlight the option called **Edit…** then press **ENTER**)
2. Press the **2$^{\text{nd}}$** key then press the **Y** = key
3. Highlight **Plot1** and press **ENTER**
4. Highlight **On** and press **ENTER**
5. Highlight the plot type at the far right and press **ENTER**
6. Press **GRAPH**

TI-Nspire

1. Enter the data into a data list (In order to access data lists select the spreadsheet option and press **enter)**

Note: Be sure to title the list by selecting the top row of the column and typing a title.

2. Press **menu** and select **3:Data** then **9:QuickGraph** then press **enter**
3. Press **menu** and select **1:Plot Type** then select **4:Normal Probability Plot** and press **enter**

Cumulative Review Exercises CR7.1–CR7.17 ● Data set available online

CR7.1 The paper **"The Psychology of Security"** (***Communications of the ACM*** **[2008]: 36–41)** states that people are more likely to gamble for a loss than to accept a guaranteed loss. For example, the authors indicate that when presented with a scenario that allows people to choose between a guaranteed gain of \$5 and a gain of \$10 if the flip of a coin results in a head, more people will choose the guaranteed gain. However, when presented with a scenario that involves choosing between a guaranteed loss of \$5 and a loss of \$10 if the flip of a coin results in a head, more people will choose the gambling option. Describe how you would design an experiment that would allow you to test this theory.

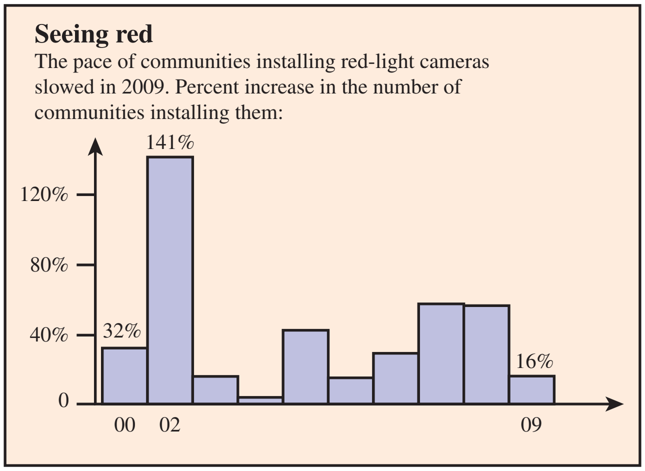
CR7.2 ● The article **"Water Consumption in the Bay Area"** (***San Jose Mercury News,*** **January 16, 2010)** reported the accompanying values for water consumption (gallons per person per day) for residential customers of 27 San Francisco Bay Area water agencies:

49 52 53 59 59 61 69 70 70 73 79 80
85 87 89 90 92 93 93 95 96 97 114 149
208 318 334

a. Use the given water consumption values to construct a boxplot. Describe any interesting features of the boxplot. Are there any outliers in the data set?
b. Calculate the mean and standard deviation for this data set.
c. If the two largest values were deleted from the data set, would the standard deviation of this new data set be greater than or less than the standard deviation you calculated for the entire data set?
d. How do the values of the mean and median of the entire data set compare? Is this consistent with the shape of the boxplot from Part (a)? Explain.

Copyright 2025 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s). Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

CR7.3 Red-light cameras are used in many places to deter drivers from running red lights. The following graphical display is similar to one that appeared in the article **"Communities Put a Halt to Red-Light Cameras" (*USA TODAY,* January 18, 2010)**. Based on this graph, would it be correct to conclude that there were fewer red-light cameras in 2009 than in 2002? Explain.



Sources: Insurance Institute for Highway Safety; USA TODAY, January 18, 2010

CR7.4 The following observations were made in an article that appeared in ***USA TODAY* (January 21, 2010)**:

> Most Americans think the Census is important, and the majority say they will participate in the count this spring—although Hispanic people, younger people and the less educated are not as enthusiastic, a Pew Research Center survey found. The poll of 1504 adults found that perceptions differ: 74% of Black people and 72% of Hispanic people rate the Census "very important" vs. 57% of White people. Believing it's important doesn't necessarily mean participating: Fewer than half of Hispanic people, 57% of Black people, and 61% of White people say they will fill out and mail the forms. Age is the greatest factor in participation: Only 36% of respondents who are 30 or younger say they will definitely participate.

Suppose that it is reasonable to regard the survey participants as representative of adult Americans.

a. Suppose that an adult American is to be selected at random and consider the following events:

I = the event that the selected individual thinks that the Census is very important
W = the event that the selected individual is a White person

Are the events I and W independent or dependent events? Explain.

b. Give an example of two census-related events different than the ones defined in Part (a) and indicate whether they are independent or dependent events. Explain your reasoning.

CR7.5 Beverly is building a fireplace. Beverly is using handmade decorative 8-inch bricks for the bottom row of the fireplace. Although the bricks are advertised as 8-inch bricks, the actual length of a brick is variable, and the brick lengths are approximately normally distributed with a mean of 8 inches and a standard deviation of 0.2 inches.

Unfortunately, Beverly didn't buy enough of the 8-inch bricks, and needs to get two more bricks to try to fill the remaining 15.9 inches. The two additional bricks cannot have a combined length of more than 15.9 inches and must fit in the 15.9-inch space with no more than 0.1 inch of space left over.

What is the probability that two randomly selected bricks will allow Beverly to complete the bottom row of the fireplace? (Hint: Define b_1 = length of the first brick and b_2 = length of the second brick. A variable that is a linear combination of independent variables that have normal distributions also has a normal distribution.)

CR7.6 ● Manatees are large marine mammals found along the coast of Florida. Data on the total number of manatee deaths per year reported by the **Florida Marine Research Institute** is given in the accompanying table. Construct a time series plot of these data and comment on any trends over time.

Year	Number of Manatee Deaths	Year	Number of Manatee Deaths
2001	325	2011	453
2002	305	2012	392
2003	380	2013	830
2004	276	2014	371
2005	396	2015	405
2006	417	2016	520
2007	317	2017	538
2008	337	2018	824
2009	429	2019	607
2010	766	2020	637

CR7.7 Two shipping services offer overnight delivery of parcels, and both promise delivery before 10 A.M. A mail-order catalog company ships 30% of its overnight packages using service 1 and 70% using service 2. Service 1 fails to meet the 10 A.M. delivery promise 10% of the time, whereas Service 2 fails to deliver by 10 A.M. 8% of the time.

Suppose that you made a purchase from this company and were expecting your package by 10 A.M., but it is late. Which shipping service was more likely to have been used?

CR7.8 The paper **"Examining Communication- and Media-Based Recreational Sedentary Behaviors Among Canadian Youth: Results from the COMPASS Study" (*Preventive Medicine* [2015]: 74–80)** estimated that the time spent playing video or computer games had a mean of 123.4 minutes per day and a standard deviation of 117.1 minutes per day. Based on this mean and standard deviation, explain why it is not reasonable to think that the distribution of the random variable x = time spent playing video or computer games is approximately normal.

CR7.9 A company receives light-bulbs from two different suppliers. Consider the variables x and y where

x = lifetime of a bulb from Supplier 1
y = lifetime of a bulb from Supplier 2

Five hundred bulbs from each supplier are tested, and the lifetime of each bulb (in hours) is recorded. The density histograms at the bottom of the page are constructed from these two sets of observations. Although these histograms are constructed using data from only 500 bulbs, they can be considered approximations to the corresponding probability distributions.

- **a.** Which probability distribution has the larger mean?
- **b.** Which probability distribution has the larger standard deviation?
- **c.** Assuming that the cost of the light-bulbs is the same for both suppliers, which supplier would you recommend? Explain.
- **d.** One of the two distributions pictured has a mean of approximately 1000, and the other has a mean of about 900. Which of these is the mean of the distribution for the variable x (lifetime for a bulb from Supplier 1)?
- **e.** One of the two distributions pictured has a standard deviation of approximately 100, and the other has a standard deviation of about 175. Which of these is the standard deviation of the distribution for the variable x (lifetime for a bulb from Supplier 1)?

CR7.10 According to a study conducted by a risk assessment firm (*Associated Press,* December 8, 2005), drivers residing within 1 mile of a restaurant are 30% more likely to be in an accident in a given policy year. Consider the following two events:

A = event that a driver has an accident during a policy year
R = event that a driver lives within 1 mile of a restaurant

Which of the following four probability statements is consistent with the findings of this survey? Justify your choice.

i. $P(A|R) = 0.3$ **iii.** $\dfrac{P(A|R)}{P(A|R^C)} = 0.3$

ii. $P(A|R^C) = 0.3$ **iv.** $\dfrac{P(A|R) - P(A|R^C)}{P(A|R^C)} = 0.3$

CR7.11 Suppose that a new Internet company Mumble.com requires all employees to take a drug test. Mumble.com can afford only the inexpensive drug test—the one with a 5% false-positive rate and a 10% false-negative rate. (That means that 5% of those who are not using drugs will incorrectly test positive and that 10% of those who are actually using drugs will test negative.) Suppose that 10% of those who work for Mumble.com are using the drugs for which Mumble is checking. (Hint: It may be helpful to draw a tree diagram to answer the questions that follow.)

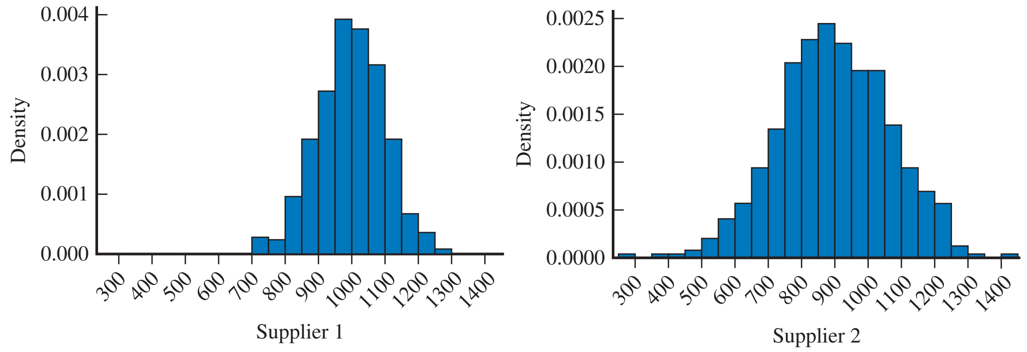
- **a.** If one employee is chosen at random, what is the probability that the employee both uses drugs and tests positive?
- **b.** If one employee is chosen at random, what is the probability that the employee does not use drugs but tests positive anyway?
- **c.** If one employee is chosen at random, what is the probability that the employee tests positive?
- **d.** If you know that a randomly chosen employee has tested positive, what is the probability that they use drugs?

CR7.12 Refer to the previous exercise. Suppose that because of the high rate of false-positives for the drug test, Mumble.com has instituted a mandatory independent second test for those who test positive on the first test.

- **a.** If one employee is selected at random, what is the probability that the selected employee uses drugs and tests positive twice?
- **b.** If one employee is selected at random, what is the probability that the employee tests positive twice?
- **c.** If you know that the randomly chosen employee has tested positive twice, what is the probability that they use drugs?
- **d.** What is the chance that an individual who does use drugs doesn't test positive twice (either this employee tests negative on the first round and doesn't need a retest, or this employee tests positive the first time and that result is followed by a negative result on the retest)?
- **e.** Discuss the benefits and drawbacks of using a retest scheme such as the one proposed in this question.

CR7.13 A chemical supply company currently has in stock 100 pounds of a certain chemical, which it sells to customers in 5-pound lots. Let x = the number of lots ordered by a

Histograms for Exercise CR 7.9



Copyright 2025 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s). Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

randomly chosen customer. The probability distribution of x is as follows:

x	1	2	3	4
$p(x)$	0.2	0.4	0.3	0.1

a. Calculate the mean value of x.
b. Calculate the variance and standard deviation of x.

CR7.14 Return to the previous exercise, and let y denote the amount of material (in pounds) left after the next customer's order is shipped. Find the mean and variance of y. (Hint: y is a linear function of x.)

CR7.15 Consider the variable

$x =$ time required for a college student to complete a standardized exam

Suppose that for the population of students at a particular university, the distribution of x is well approximated by a normal distribution with mean 45 minutes and standard deviation 5 minutes.

a. If 50 minutes is allowed for the exam, what proportion of students at this university would be unable to finish in the allotted time?
b. How much time should be allowed for the exam if you wanted 90% of the students taking the test to be able to finish in the allotted time?
c. How much time is required for the fastest 25% of all students to complete the exam?

CR7.16 ● The accompanying data on $x =$ student-teacher ratio for a random sample of 25 high schools in Maine selected from a population of 85 high schools are consistent with summary values for the state of Maine that appeared in an article in the ***Bangor Daily News* (September 22, 2016)**. The corresponding normal scores are also shown.

Student-Teacher Ratio (x)	Normal Score
9.0	−1.868
10.0	−1.403
11.0	−1.128
11.2	−0.919
11.6	−0.744
11.7	−0.589
11.8	−0.448
11.9	−0.315
12.0	−0.187
12.1	−0.062
12.5	0.062
12.6	0.187
13.0	0.315
13.2	0.448
13.6	0.589
13.7	0.744
14.0	0.919
14.5	1.128
14.9	1.403
15.0	1.868

(continued)

Construct a normal probability plot. Is it reasonable to assume that the distribution of student-teacher ratio is approximately normal?

CR7.17 ● The following data are a sample of survival times (days from diagnosis) for patients suffering from chronic leukemia of a certain type:

7	47	58	74	177	232	273	285
317	429	440	445	455	468	495	497
532	571	579	581	650	702	715	779
881	900	930	968	1077	1109	1314	1334
1367	1534	1712	1784	1877	1886	2045	2056
2260	2429	2509					

a. Construct a relative frequency distribution for this data set, and draw the corresponding histogram.
b. Would you describe this histogram as having a positive or a negative skew?
c. Would you recommend transforming the data? Explain.

Copyright 2025 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s). Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

Chapter 8

Sampling Variability and Sampling Distributions

Alena Haurylik/Shutterstock.com

The inferential methods that you will see in Chapters 9–15 use sample data to learn about population characteristics. For example, suppose that μ denotes the true mean fat content of quarter-pound hamburgers marketed by a national fast-food chain. To learn something about μ, you might select a sample of $n = 50$ hamburgers and determine the fat content for each one. The sample data might produce a mean of $\bar{x} = 28.4$ grams. How close is this sample mean to the population mean, μ? If you select another sample of 50 quarter-pound burgers and then determine the sample mean fat content, would this second value be near 28.4, or might it be quite different?

These questions can be addressed by studying what is called the *sampling distribution* of $\bar{x}$. Just as the probability distribution of a numerical variable describes its long-run behavior, the sampling distribution of $\bar{x}$ provides information about the long-run behavior of $\bar{x}$ when samples are selected from a population.

In this chapter, the sampling distribution of a sample proportion (the fraction of individuals or objects in a sample that have some characteristic of interest) is also considered. The sampling distribution of a sample proportion, $\hat{p}$, provides information about the long-run behavior of the sample proportion. This knowledge allows you to make inferences about a population proportion.

Learning Objectives

After completing this chapter, you will be able to:

- Explain why there is sample-to-sample variability in the value of a sample statistic.
- Explain how the sampling distribution of a sample statistic describes sample-to-sample variability in the values of a statistic.
- Describe the relationship between the standard deviations of the sampling distributions of $\bar{x}$ and $\hat{p}$ and sample size.
- Describe general properties of the sampling distribution of $\bar{x}$.
- Describe general properties of the sampling distribution of $\hat{p}$.
- Determine when the sampling distribution of $\bar{x}$ or $\hat{p}$ is approximately normal.
- Reason informally about the value of a population mean using properties of the sampling distribution of the sample mean.
- Reason informally about the value of a population proportion using properties of the sampling distribution of the sample proportion.

Section 8.1 Statistics and Sampling Variability

A number calculated from the values in a sample is called a **statistic**. Values of statistics such as the sample mean $\bar{x}$, the sample median, the sample standard deviation s, or the proportion of individuals in a sample that possess a particular property $\hat{p}$, provide information about population characteristics.

It is common to learn about the value of a population characteristic by selecting a sample from the population. For example, to learn about the mean credit card balance for students at a university, you might select a sample of 50 students. Each student could be asked about their credit card balance, resulting in a value of x = current balance.

If you were to construct a histogram of the 50 sample x values, you could view this histogram as a rough approximation of the population distribution of x. In a similar way, you can view the sample mean $\bar{x}$ (the mean of a sample) as an estimate of μ, the mean of the population distribution.

It would be nice if the value of $\bar{x}$ was equal to the value of the population mean μ, but this is not usually the case. Not only will the value of $\bar{x}$ for a particular sample from a population usually differ from μ, but the $\bar{x}$ values from different samples also usually differ from one another. For example, two different samples of 50 student credit card balances will usually result in different $\bar{x}$ values. This sample-to-sample variability must be considered when generalizing from a sample to the population from which it was selected.

Definitions

Statistic: A number calculated from values in a sample.

Sampling variability: The observed value of a statistic depends on the particular sample selected from the population and it will vary from sample to sample. This variability is called **sampling variability**.

Example 8.1 Exploring Sampling Variability

● Consider a small population consisting of the 20 students enrolled in a class. The amount of money (in dollars) each of the 20 students spent on textbooks for the current semester is shown in the following table:

Student	Amount Spent on Books	Student	Amount Spent on Books	Student	Amount Spent on Books
1	367	8	370	15	433
2	358	9	378	16	284
3	442	10	268	17	331
4	361	11	419	18	259
5	375	12	363	19	330
6	395	13	365	20	423
7	322	14	362		

For this population,

$$\mu = \frac{367 + 358 + \cdots + 423}{20} = 360.25$$

Suppose you don't know the value of the population mean, so you decide to estimate μ by taking a random sample of five students and calculating $\bar{x}$, the sample mean amount spent on textbooks. Is this a reasonable thing to do? Is the estimate that results likely to be close to the value of μ, the population mean?

To answer these questions, consider a simple experiment that allows you to study the behavior of the statistic $\bar{x}$ when random samples of size 5 are repeatedly selected. Begin by selecting a random sample of size 5 from this population. This can be done by writing

● **Data set available online**

the numbers from 1 to 20 on otherwise identical slips of paper, mixing them well, and then selecting 5 slips without replacement. The numbers on the slips selected identify which of the 20 students will be included in the sample. Alternatively, either a table of random digits or a random number generator could be used to determine which 5 students should be selected.

Minitab was used to obtain five random numbers between 1 and 20, resulting in 17, 20, 7, 11, and 9. This results in the following sample of amounts spent on books:

331 423 322 419 378

For this sample,

$$\bar{x} = \frac{1873}{5} = 374.60$$

The sample mean is larger than the population mean of $360.25 by about $14. Is this difference typical, or is this particular sample mean unusually far away from μ? Taking more samples will provide additional insight.

Four more random samples (Samples 2–5) from this same population are shown here.

Sample 2		**Sample 3**		**Sample 4**		**Sample 5**	
Student	x	**Student**	x	**Student**	x	**Student**	x
4	361	15	433	20	423	18	259
15	433	12	363	16	284	8	370
12	363	3	442	19	330	9	378
1	367	7	322	1	367	7	322
18	259	18	259	8	370	14	362
$\bar{x}$	**356.60**	$\bar{x}$	**363.80**	$\bar{x}$	**354.80**	$\bar{x}$	**338.20**

Because $\mu = 360.25$ for this population, you can see the following:

1. The value of $\bar{x}$ varies from one random sample to another (sampling variability).
2. Some samples produced $\bar{x}$ values greater than μ (Samples 1 and 3), whereas others produced values less than μ (Samples 2, 4, and 5).
3. Samples 2, 3, and 4 produced $\bar{x}$ values that were fairly close to the population mean, but Sample 5 resulted in a value that was $22 less than the population mean.

Continuing with the experiment, 45 additional random samples of size $n = 5$ were selected. The resulting sample means are as follows:

Sample	$\bar{x}$	**Sample**	$\bar{x}$	**Sample**	$\bar{x}$
6	374.6	21	355.0	36	353.4
7	356.6	22	407.2	37	379.6
8	363.8	23	380.0	38	352.6
9	354.8	24	377.4	39	342.2
10	338.2	25	341.2	40	362.6
11	375.6	26	316.0	41	315.4
12	379.2	27	370.0	42	366.2
13	341.6	28	401.0	43	361.4
14	355.4	29	347.0	44	375.0
15	363.8	30	373.8	45	401.4
16	339.6	31	382.8	46	337.0
17	348.2	32	320.4	47	387.4
18	430.8	33	313.6	48	349.2
19	388.8	34	387.6	49	336.8
20	352.8	35	314.8	50	364.6

Copyright 2025 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s). Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

Figure 8.1 is a density histogram of the 50 sample means. It provides insight about the behavior of $\bar{x}$. Most samples resulted in $\bar{x}$ values that are reasonably near $\mu = 360.25$, falling between 335 and 395. A few samples, however, produced values that were far from μ. If you were to take a sample of size 5 from this population and use $\bar{x}$ as an estimate of the population mean μ, you should *not* necessarily expect $\bar{x}$ to be close to μ.

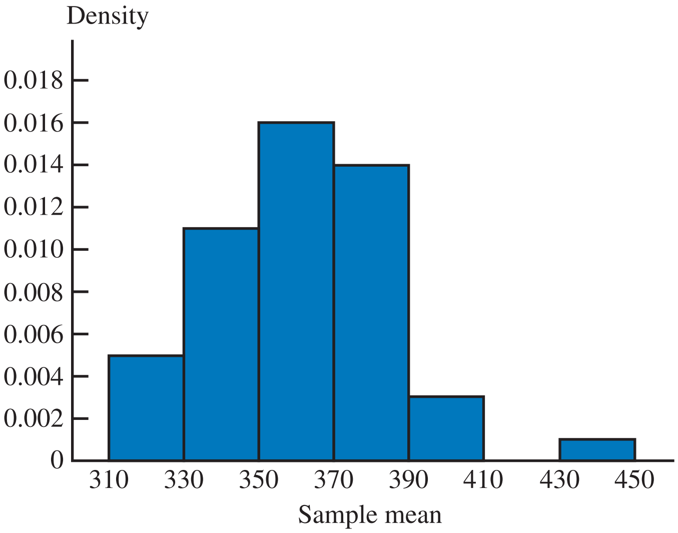


Figure 8.1
Density histogram of $\bar{x}$ values from the 50 random samples of Example 8.1.

The density histogram in Figure 8.1 visually conveys information about the sampling variability in the statistic $\bar{x}$. It provides an approximation to the distribution of $\bar{x}$ values that would have been observed if you had considered every different possible sample of size 5 from this population.

In the example just considered, the approximate sampling distribution of the statistic $\bar{x}$ was obtained by considering just 50 different samples. The actual **sampling distribution** results from considering *all* possible samples of size n.

Definition

Sampling distribution: The distribution that would be formed by considering the value of a sample statistic for every possible sample of a given size from a population.

The sampling distribution of a statistic, such as $\bar{x}$, provides important information about variability in the values of the statistic. The density histogram of Figure 8.1 is an *approximation* of the sampling distribution of the statistic $\bar{x}$ for samples of size 5 from the population described in Example 8.1.

The actual sampling distribution of $\bar{x}$ could have been determined by considering every possible sample of size 5 from the population of 20 students, calculating the mean for each sample, and then constructing a density histogram of the $\bar{x}$ values. But this would have been a lot of work—there are 15,504 different possible samples of size 5. And, for more realistic cases with larger population and larger sample sizes, it becomes even more difficult because there are so many possible samples that must be considered.

Fortunately, as you look at more examples in the sections that follow, patterns emerge that will enable you to describe some important aspects of the sampling distributions for some statistics without actually having to look at all possible samples.

Exercises 8.1–8.9

8.1 Explain the difference between a population characteristic and a statistic.

8.2 A random sample of 1000 students at a large college included 428 who had more than one credit cards. For this sample, $\hat{p} = \frac{428}{1000} = 0.428$. If another random sample of 1000 students from this university were selected, would you expect that $\hat{p}$ for that sample would also be 0.428? Explain why or why not.

8.3 For each of the following statements, identify the number that appears in boldface type as the value of either a population characteristic or a statistic:

- **a.** A department store reports that **84%** of all customers who use the store's credit plan pay their bills on time.
- **b.** A sample of 100 students at a large university has a mean age of **24.1** years.
- **c.** The Department of Motor Vehicles reports that **22%** of all vehicles registered in a particular state are imports.
- **d.** A hospital reports that based on the 10 most recent cases, the mean length of stay for surgical patients is **6.4** days.
- **e.** A consumer group, after testing 100 batteries of a certain brand, reported an average life of **63** hours of use.

8.4 Consider a population consisting of the following five values, which represent the number of video downloads during the academic year for each of five housemates:

8 14 16 10 11

- **a.** Calculate the mean of this population.
- **b.** Select a random sample of size 2 by writing the five numbers in this population on slips of paper, mixing them, and then selecting two. Calculate the mean for your sample.
- **c.** Repeatedly select random samples of size 2, and calculate the $\bar{x}$ value for each sample until you have the $\bar{x}$ values for 25 samples.
- **d.** Construct a density histogram using the 25 $\bar{x}$ values. Are most of the $\bar{x}$ values near the population mean? Do the $\bar{x}$ values differ a lot from sample to sample, or do they tend to be similar? (Hint: See Example 8.1.)

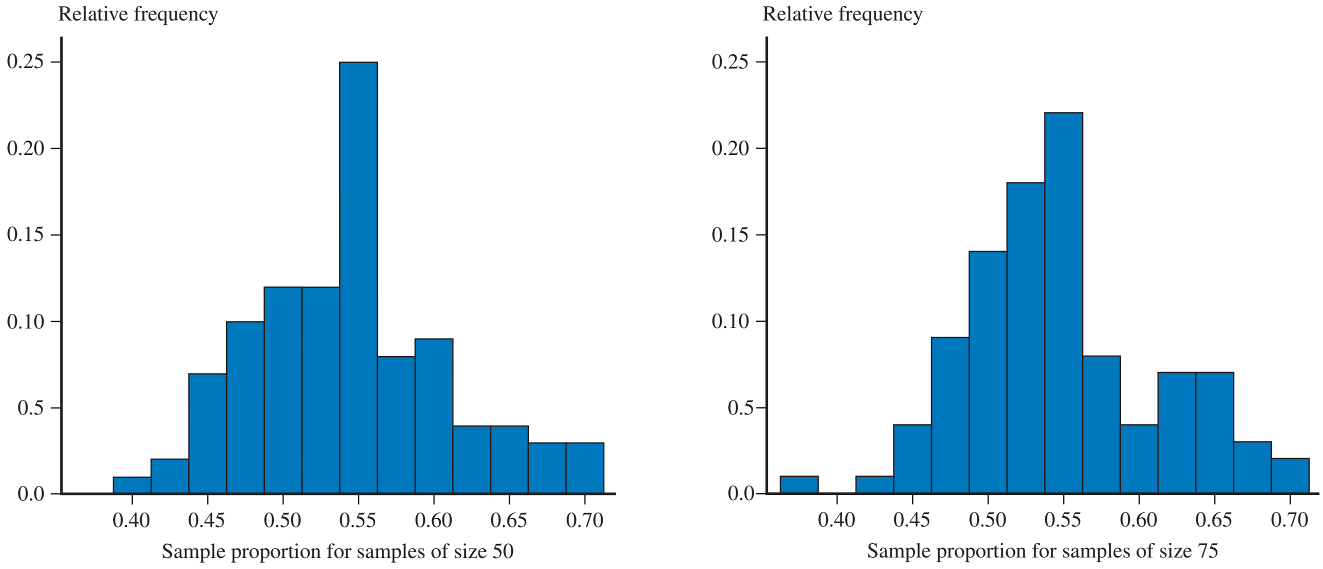
8.5 Consider the two relative frequency histograms at the bottom of this page. The histogram on the left was constructed by selecting 100 different random samples of size 50 from a population consisting of 55% undergraduate students and 45% graduate students. For each sample, the sample proportion of undergraduate students, was calculated. The 100 sample proportions were used to construct the histogram. The histogram on the right was constructed in a similar way, except that the samples were size 75 instead of size 50.

- **a.** Which of the two histograms shows more sample-to-sample variability? How can you tell?
- **b.** For which of the two sample sizes, $n = 50$ or $n = 75$, do you think the value of a sample proportion is more likely to be close to 0.55? What about the given histograms supports your choice?

8.6 Suppose that the sampling experiment described in Example 8.1 had used samples of size 10 rather than size 5.

- **a.** If 50 samples of size 10 were selected, the $\bar{x}$ value for each sample calculated, and a density

Histograms for Exercise 8.5



Copyright 2025 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s). Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

histogram constructed, how do you think this histogram would differ from the density histogram constructed for samples of size 5 (Figure 8.1)?

b. In what way do you think the density histograms would be similar?

8.7 Consider the following population: {1, 2, 3, 4}. For this population the mean is

$$\mu = \frac{1 + 2 + 3 + 4}{4} = 2.5$$

Suppose that a random sample of size 2 is to be selected without replacement from this population. There are 12 possible samples (assuming that the order in which observations are selected is taken into account):

1, 2 1, 3 1, 4 2, 1 2, 3 2, 4
3, 1 3, 2 3, 4 4, 1 4, 2 4, 3

a. Calculate the sample mean for each of the 12 possible samples.

b. Use the sample means to construct the sampling distribution of $\bar{x}$. Display the sampling distribution as a density histogram.

c. Suppose that a random sample of size 2 is to be selected, but this time sampling will be done with replacement. Using a method similar to that of Part (a), construct the sampling distribution of $\bar{x}$. (Hint: There are 16 different possible samples in this case.)

d. In what ways are the two sampling distributions of Parts (b) and (c) similar? In what ways are they different?

8.8 Simulate sampling from the population of Exercise 8.7 by using four slips of paper individually marked 1, 2, 3, and 4. Select a sample of size 2 without replacement, and calculate $\bar{x}$. Repeat this process 50 times, and construct a density histogram of the 50 $\bar{x}$. values. How does this sampling distribution compare to the actual sampling distribution of $\bar{x}$. from Exercise 8.7, Part (b)?

8.9 Consider the following population: {2, 3, 3, 4, 4}. The value of μ is 3.2, but suppose that this is not known to an investigator. Three possible statistics for estimating μ are

Statistic 1: the sample mean, $\bar{x}$
Statistic 2: the sample median
Statistic 3: the average of the largest and the smallest values in the sample

A random sample of size 3 will be selected without replacement. If you disregard the order in which the observations are selected, there are 10 possible samples that might result (writing 3 and 3*, 4 and 4* to distinguish the two 3's and the two 4's in the population):

2, 3, 3* 2, 3, 4 2, 3, 4* 2, 3*, 4 2, 3*, 4*
2, 4, 4* 3, 3*, 4 3, 3*, 4* 3, 4, 4* 3*, 4, 4*

a. For each of these 10 samples, calculate the values of Statistics 1, 2, and 3.

b. Construct the sampling distribution of each of these statistics.

c. Which statistic would you recommend for estimating μ? Explain your reasoning.

Section 8.2 The Sampling Distribution of a Sample Mean

When you want to learn about the value of a population mean μ, it is natural to consider the sample mean $\bar{x}$. To understand how inferential procedures based on $\bar{x}$ work, begin by looking at how $\bar{x}$ varies in value from one sample to another.

The behavior of $\bar{x}$ is described by its sampling distribution. The sample size n and characteristics of the population—its shape, mean value μ, and standard deviation σ—are important in determining properties of the sampling distribution of $\bar{x}$.

It is helpful first to consider the results of some sampling experiments. In Examples 8.2 and 8.3, a known population distribution is specified and a sample size n is selected. Then 500 different samples of size n are selected, the value of $\bar{x}$ for each sample is calculated, and a density histogram of these 500 $\bar{x}$ values is constructed. Because 500 is reasonably large (a reasonably long sequence of samples), the density histogram of the $\bar{x}$ values should closely resemble the actual sampling distribution of $\bar{x}$ (which would be obtained by considering all possible samples).

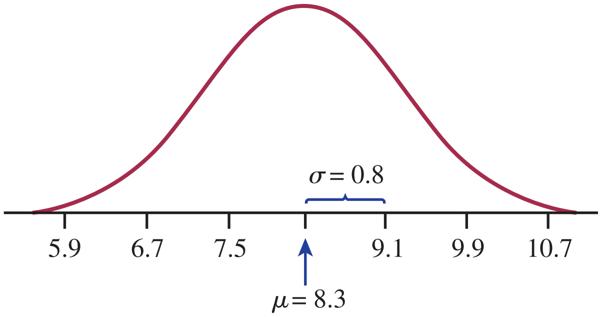
This experiment is repeated for several different values of n to see how the choice of sample size affects the sampling distribution. This will enable you to identify some patterns that will be helpful in understanding important properties of the sampling distribution of $\bar{x}$.

Example 8.2 Blood Platelet Volume

The paper **"Mean Platelet Volume Could Be Possible Biomarker in Early Diagnosis and Monitoring of Gastric Cancer" (*Platelets* [2014]: 592–594)** includes data that suggest that the distribution of platelet volume for patients who have gastric cancer is approximately normal with mean $\mu = 8.3$ and standard deviation $\sigma = 0.8$.

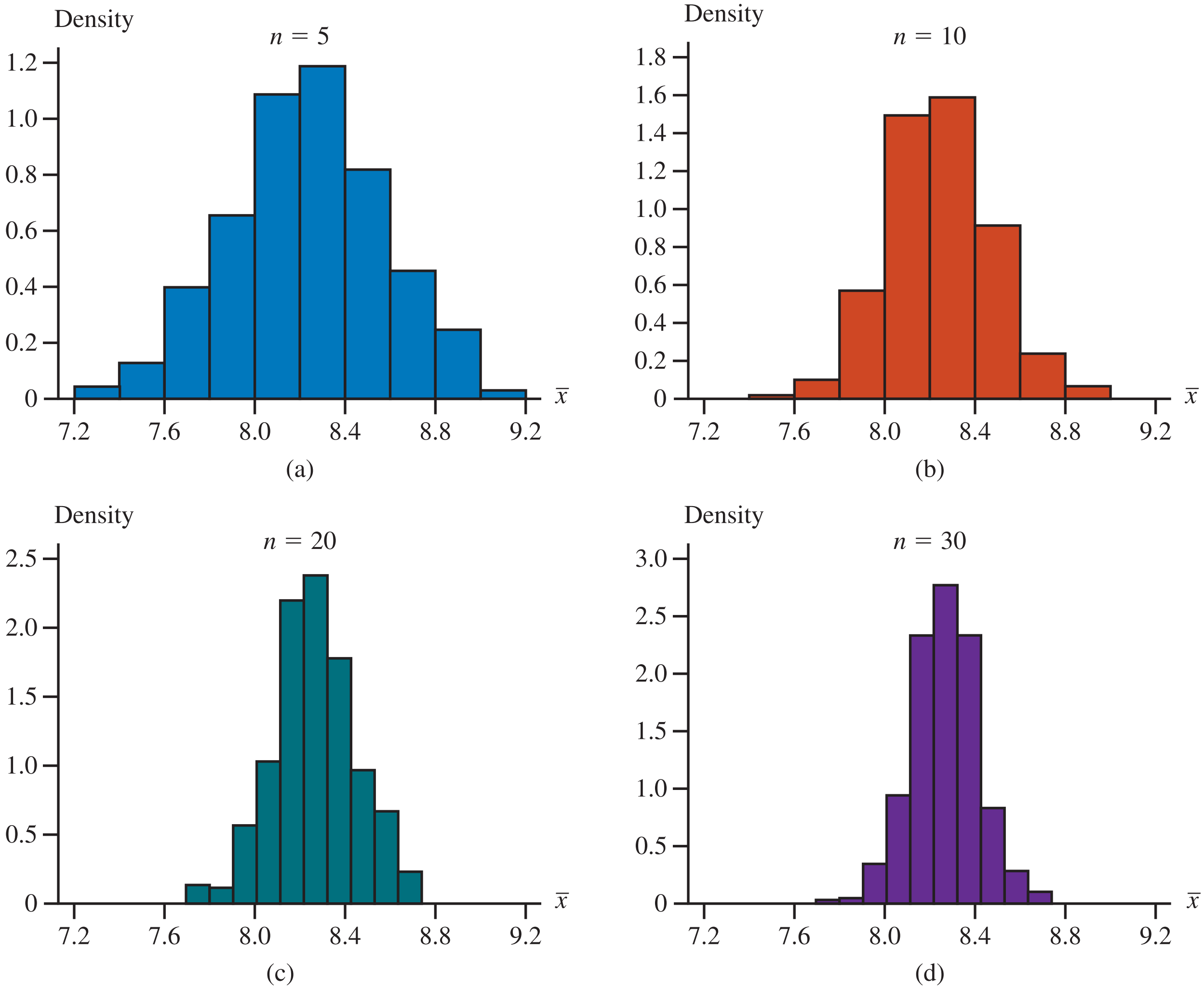
Figure 8.2 shows a normal curve centered at 8.3, the mean value of platelet volume. The value of the population standard deviation, 0.8, determines the extent to which the *x* distribution spreads out about its mean value.

Figure 8.2
Normal distribution of platelet size ($\mu = 8.3$ and $\sigma = 0.8$).



Minitab was used to select 500 random samples from this population distribution, with each sample consisting of $n = 5$ observations. A density histogram of the resulting 500 $\bar{x}$ values appears in Figure 8.3(a). This procedure was repeated for samples of size $n = 10$, $n = 20$, and $n = 30$. The resulting density histograms of the $\bar{x}$ values are displayed in Figure 8.3(b)–(d).

Figure 8.3
Density histograms for $\bar{x}$ based on 500 samples, each consisting of n observations, for Example 8.2: (a) $n = 5$; (b) $n = 10$; (c) $n = 20$; (d) $n = 30$.



Copyright 2025 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s). Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

Notice the shapes of the histograms. Each of the four histograms is approximately normal in shape. The resemblance would be even more striking if each histogram had been based on many more than 500 $\bar{x}$ values.

Second, notice that each histogram is centered approximately at 8.3, the mean of the population being sampled. Had the histograms been constructed using $\bar{x}$ values from every possible sample, the histogram centers would have been exactly equal to the population mean, 8.3.

What Do You Think?

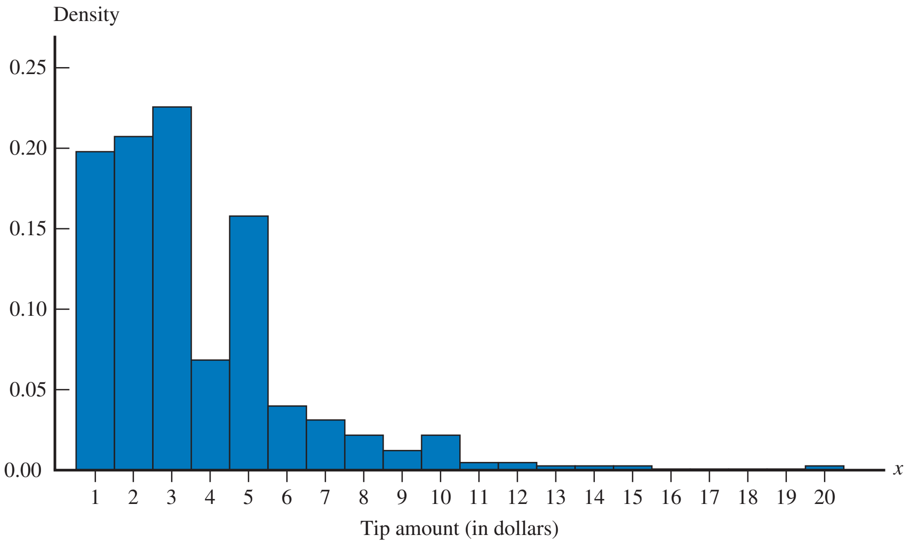
Consider the sample-to-sample variability reflected in each of the four histograms in Figure 8.3. What do you notice about how the variability changes as the sample size increases?

The final aspect of the histograms to note is their variability relative to one another. The smaller the value of n, the greater the extent to which the sampling distribution spreads out around the population mean value. This is why the histograms for $n = 20$ and $n = 30$ are based on narrower class intervals than those for the two smaller sample sizes. For the larger sample sizes, most of the $\bar{x}$ values are quite close to 8.3. This is the effect of averaging. When n is small, a single unusual x value can result in an $\bar{x}$ value far from the center. With a larger sample size, any unusual x values, when averaged with the other sample values, still tend to yield an $\bar{x}$ value close to μ. Combining these insights yields a result that should appeal to your intuition: $\bar{x}$ values based on a large sample will tend to be closer to μ than $\bar{x}$ values based on a small sample.

Example 8.3 Tipping Uber Drivers

In this example, you will consider properties of the $\bar{x}$ distribution when the population is quite skewed. The City of Chicago began requiring rideshare companies to report data to the city in 2018, and makes the data available to the public through the **Chicago Data Portal**. One of the variables reported is the amount of the tip given for each Uber ride. Figure 8.4 displays a density histogram of the tip amounts (in dollars) for 425 Uber rides for which the rider gave a tip (not all Uber riders leave a tip for the driver). The histogram has a long upper tail, indicating that most tips were $5.00 or less but that a small percentage of the tips were much greater.

Figure 8.4
The population distribution for Example 8.3 (μ = **$3.55).**

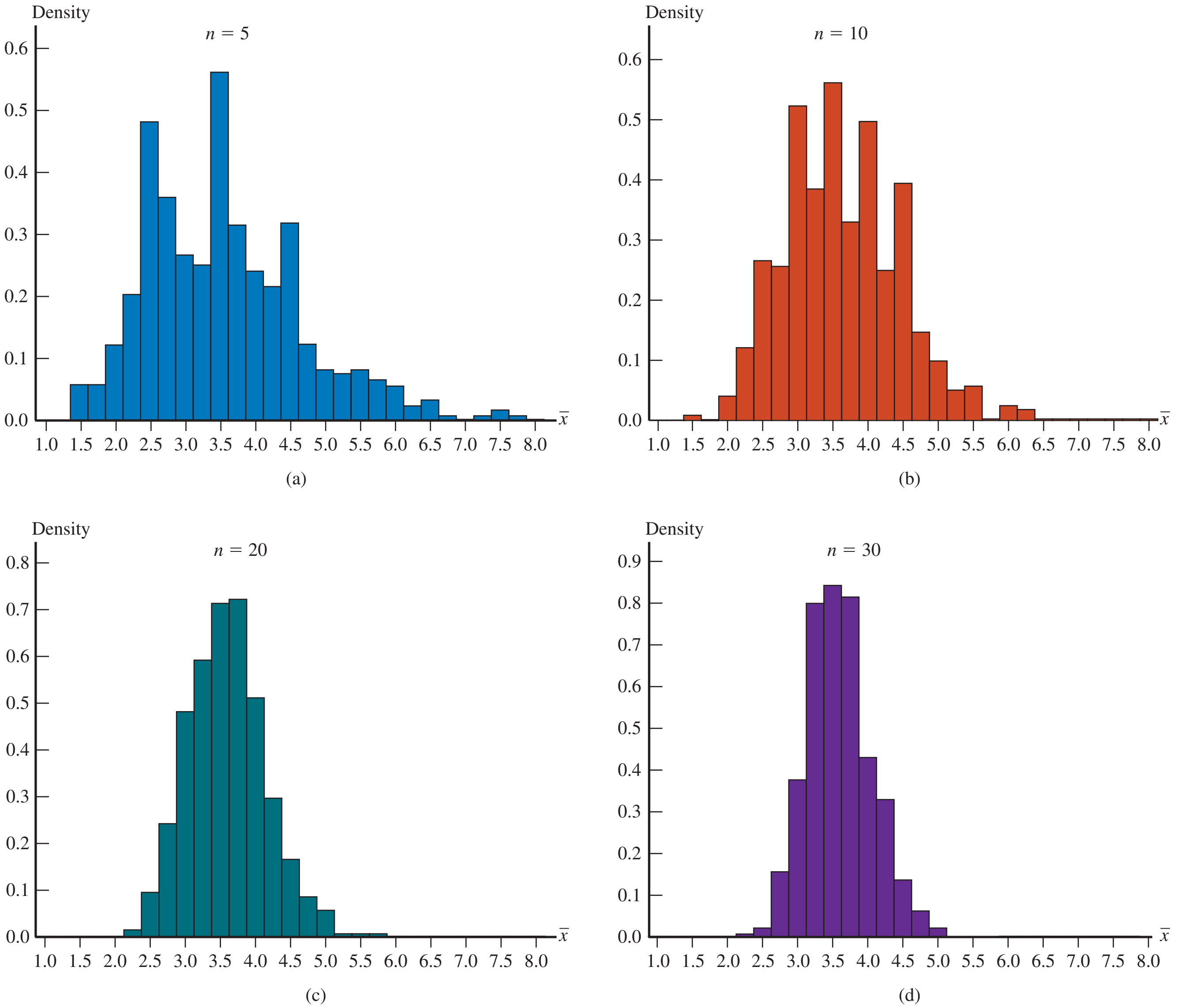


Copyright 2025 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s). Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

If you think of these 425 tip amounts as the population, the histogram in Figure 8.4 displays the population distribution. The skewed shape makes identification of the mean value from the histogram more difficult. The mean of the 425 values in the population was calculated to be $\mu = \$3.55$. The median value for the population is \$3.00, which is less than μ because the distribution is positively skewed.

For each of the sample sizes $n = 5$, 10, 20, and 30, 500 random samples of size n were selected. This was done with replacement to approximate more nearly the usual situation, in which the sample size n is only a small fraction of the population size. Density histograms of the 500 $\bar{x}$ values were constructed for each of the four sample sizes. These histograms are displayed in Figure 8.5.

Figure 8.5
Four density histograms of 500 $\bar{x}$ values for Example 8.3:
(a) $n = 5$; (b) $n = 10$; (c) $n = 20$; (d) $n = 30$.



(a) (b) (c) (d)

As with the samples from the normal population of Example 8.2, the means of the 500 $\bar{x}$ values for the four different sample sizes are all close to the population mean $\mu = \$3.55$. If each histogram had been constructed using all possible samples rather than just 500 of them, they would be centered at exactly \$3.55. Comparison of the four $\bar{x}$ histograms in Figure 8.5 also shows that as n increases, the histogram's spread about its center decreases.

Copyright 2025 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s). Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

This was also true for the normal population of the previous example. There is less sample-to-sample variability in $\bar{x}$ for large samples than for small samples.

One aspect of the histograms in Figure 8.5 distinguishes them from those based on a random sample from a normal population (Figure 8.3). The histograms in Figure 8.5 are skewed and differ in shape more, but they become progressively more symmetric as the sample size increases. You can also see that for $n = 30$, the histogram has a shape much like a normal distribution. This is the effect of averaging. Even when n is large, one of the few large x values in the population doesn't appear in the sample very often. When one does appear, its contribution to $\bar{x}$ is outweighed by the contributions of more typical sample values.

The normal shape of the histogram for $n = 30$ is what is predicted by the *Central Limit Theorem*, which will be introduced shortly. According to this theorem, even if the population distribution does not look like a normal distribution, the sampling distribution of $\bar{x}$ is approximately normal in shape when the sample size n is reasonably large.

General Properties of the Sampling Distribution of $\bar{x}$

Examples 8.2 and 8.3 suggest that for any n, the center of the $\bar{x}$ distribution (the mean value of $\bar{x}$) coincides with the mean of the population being sampled. Also notice that the variability of the $\bar{x}$ distribution decreases as n increases. This indicates that the standard deviation of $\bar{x}$ is smaller for large n than for small n. The density histograms in Figures 8.3 and 8.5 also suggest that in some cases, the $\bar{x}$ distribution is approximately normal in shape. These observations are stated more formally in the following general properties.

General Properties of the Sampling Distribution of $\bar{x}$

Notation

$\bar{x}$	Sample mean
n	Sample size
μ	Population mean
σ	Population standard deviation
$\mu_{\bar{x}}$	Mean of the sampling distribution of $\bar{x}$
$\sigma_{\bar{x}}$	Standard deviation of the sampling distribution of $\bar{x}$

When random samples are selected from a population, the following are properties of the sampling distribution of $\bar{x}$:

Property 1. $\mu_{\bar{x}} = \mu$

Property 2. $\sigma_{\bar{x}} = \dfrac{\sigma}{\sqrt{n}}$

This formula is exact if the population is infinite, and is approximately correct if the population is finite and no more than 10% of the population is included in the sample.

Property 3. When the population distribution is normal, the sampling distribution of $\bar{x}$ is also normal for any sample size n.

Property 4. (Central Limit Theorem) When n is sufficiently large, the sampling distribution of $\bar{x}$ is well approximated by a normal curve, even when the population distribution is not normal.

Property 1, $\mu_{\bar{x}} = \mu$, states that the sampling distribution of $\bar{x}$ is always centered at the mean of the population sampled.

Property 2, $\sigma_{\bar{x}} = \dfrac{\sigma}{\sqrt{n}}$, not only states that the variability of the sampling distribution of $\bar{x}$ decreases as n increases, but also gives a precise relationship between the standard

Copyright 2025 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s). Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

deviation of the $\bar{x}$ distribution and the population standard deviation and sample size. For example, when $n = 4$,

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}} = \frac{\sigma}{\sqrt{4}} = \frac{\sigma}{2}$$

so the $\bar{x}$ distribution has a standard deviation only half as large as the population standard deviation.

Properties 3 and 4 specify when the $\bar{x}$ distribution is normal (when the population is normal) or approximately normal (when the sample size is large).

Figure 8.6 illustrates these properties by showing several $\bar{x}$ distributions superimposed over a graph of the population distribution.

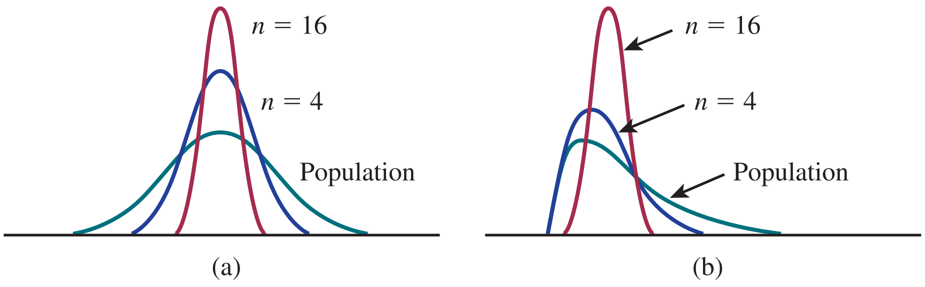


Figure 8.6 Population distribution and sampling distributions of $\bar{x}$:
(a) symmetric population;
(b) skewed population.

The Central Limit Theorem, Property 4, states that when n is sufficiently large, the sampling distribution of $\bar{x}$ is approximately normal for any population distribution. This result has enabled statisticians to develop procedures for drawing conclusions about a population mean μ using large samples, even when the shape of the population distribution is unknown.

Recall that a variable is standardized by subtracting the mean value and then dividing by its standard deviation. Using Properties 1 and 2 to standardize $\bar{x}$ gives an important consequence of the last two properties.

If n is large or the population distribution is normal, the standardized variable

$$z = \frac{\bar{x} - \mu_{\bar{x}}}{\sigma_{\bar{x}}} = \frac{\bar{x} - \mu}{\frac{\sigma}{\sqrt{n}}}$$

has (at least approximately) a standard normal (z) distribution.

Applying the Central Limit Theorem requires a rule of thumb for deciding when n is large enough. Look back at Figure 8.5, which shows the approximate sampling distribution of $\bar{x}$ for $n = 5$, 10, 20, and 30 when the population distribution is quite skewed. The histogram for $n = 5$ is not well described by a normal curve, and this is still true of the histogram for $n = 10$, particularly in the tails of the histogram. Among the four histograms, only the histogram for $n = 30$ has a shape that is reasonably well described by a normal curve.

On the other hand, when the population distribution is normal, the sampling distribution of $\bar{x}$ is normal for any n. If the population distribution is somewhat skewed but not to the extent of Figure 8.4, you might expect the sampling distribution of $\bar{x}$ to be a bit skewed for $n = 5$ but quite well described by a normal curve for n as small as 10 or 15.

How large n must be for the sampling distribution of $\bar{x}$ to be approximately normal depends on how much the population distribution differs from a normal distribution. The closer the population distribution is to a normal distribution, the smaller the value of n necessary for the Central Limit Theorem approximation to be accurate.

Copyright 2025 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s). Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

Many statisticians recommend the following conservative rule:

> The Central Limit Theorem can safely be applied if n is greater than or equal to 30.

Example 8.4 Courting Scorpion Flies

Understand the context ➤

The authors of the paper **"Should I Stay or Should I Go? Condition- and Status-Dependent Courtship Decisions in the Scorpion Fly Panorpa Cognate" (*Animal Behaviour* [2009]: 491–497)** studied the courtship behavior of mating scorpion flies. One variable of interest was x = courtship time, which was defined as the time from the beginning of a female-male interaction until mating.

Data from the paper suggest that it is reasonable to think that the mean and standard deviation of x are $\mu = 117.1$ minutes and $\sigma = 109.1$ minutes. Notice that the distribution of courtship times cannot be normal because for a normal distribution centered at 117.1 and with such a large standard deviation, it would not be uncommon to observe negative values, but courtship time can't have a negative value.

The sampling distribution of

$\bar{x}$ = mean courtship time for a random sample of 20 scorpion fly mating pairs

would have mean

$$\mu_{\bar{x}} = \mu = 117.1 \text{ minutes}$$

This tells you that the sampling distribution is centered at 117.1. The standard deviation of $\bar{x}$ is

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}} = \frac{109.1}{\sqrt{20}} = 24.40$$

which is smaller than the population standard deviation σ. Because the population distribution is not normal and because the sample size is smaller than 30, it is not reasonable to assume that the sampling distribution of $\bar{x}$ is approximately normal in shape.

What Do You Think?

Suppose that the sample size in Example 8.4 had been 100 rather than 20.

- Would the mean of the sampling distribution still be 117.1?
- Would the standard deviation of the sampling distribution still be 24.40?
- Would it be reasonable to think that the sampling distribution of the sample mean was approximately normal?

Example 8.5 Soft-Drink Volumes

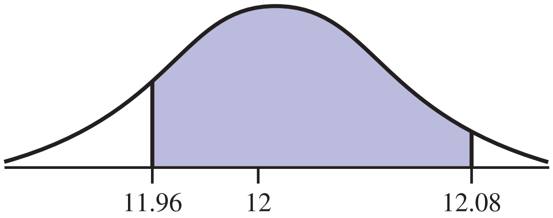
Understand the context ➤

A soft-drink manufacturer claims that, on average, cans contain 12 ounces of soda. Let x denote the actual volume of soda in a randomly selected can. Suppose that x is normally distributed with $\sigma = 0.16$ ounces. Sixteen cans are to be selected, and the soda volume will be determined for each one. Let $\bar{x}$ denote the resulting sample mean soda volume. Because the x distribution is normal, the sampling distribution of $\bar{x}$ is also normal. *If the manufacturer's claim is correct,* the sampling distribution of $\bar{x}$ has a mean value of $\mu_{\bar{x}} = \mu = 12$ and a standard deviation of

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}} = \frac{0.16}{\sqrt{16}} = 0.04$$

Copyright 2025 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s). Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

Formulate a plan ➤ To calculate a probability involving $\bar{x}$, you can use what you know about normal distributions. For example, the probability that the sample mean soda volume is between 11.96 ounces and 12.08 ounces is the area between 11.96 and 12.08 under the normal curve with mean 12 and standard deviation 0.04, as shown in the following figure.



This area can be found using technology or Appendix Table 2:

Do the work ➤

$$\begin{aligned} P(11.96 \le \bar{x} \le 12.08) &= \text{area under the } z \text{ curve between } -1.0 \text{ and } 2.0 \\ &= 0.8185 \end{aligned}$$

The probability that the sample mean soda volume is at most 11.9 ounces is

$$P(\bar{x} \le 11.9) = (\text{area under the } z \text{ curve to the left of } -2.5) = 0.0062$$

Interpret the results ➤ If the x distribution is as described and the manufacturer's claim is correct, a sample mean soda volume based on a random sample of 16 observations is less than 11.9 ounces for fewer than 1% of all such samples. If you observed an $\bar{x}$ value that is smaller than 11.9 ounces, this would cast doubt on the manufacturer's claim that the average soda volume is 12 ounces.

Example 8.6 Fat Content of Hot Dogs

Understand the context ➤ A hot dog manufacturer claims that one of its brands of hot dogs has an average fat content of $\mu = 18$ grams per hot dog. Consumers of this brand would probably not be disturbed if the mean is less than 18 but would be unhappy if it is greater than 18. Let x denote the fat content of a randomly selected hot dog, and suppose that σ, the standard deviation of the x distribution, is 1.

An independent testing organization is asked to analyze a random sample of 36 hot dogs. Let $\bar{x}$ be the average fat content for this sample. The sample size, $n = 36$, is large enough to rely on the Central Limit Theorem and so it is reasonable to regard the $\bar{x}$ distribution as approximately normal. The standard deviation of the $\bar{x}$ distribution is

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}} = \frac{1}{\sqrt{36}} = 0.167$$

If the manufacturer's claim is correct, you know that $\mu_{\bar{x}} = \mu = 18$ grams.

Suppose that the sample resulted in a mean of $\bar{x} = 18.4$ grams. Does this result suggest that the manufacturer's claim is incorrect?

Formulate a plan ➤ You can answer this question by looking at the sampling distribution of $\bar{x}$. Because of sampling variability, even if $\mu = 18$, you know that $\bar{x}$ will not usually be exactly 18. But, is it likely that you would see a sample mean at least as great as 18.4 when the population mean is really 18?

If the company's claim is correct,

Do the work ➤

$$\begin{aligned} P(\bar{x} \ge 18.4) &\approx P\left(z \ge \frac{18.4 - 18}{0.167}\right) \\ &= P(z \ge 2.40) \\ &= \text{area under the } z \text{ curve to the right of } 2.40 \\ &= 0.0082 \end{aligned}$$

Copyright 2025 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s). Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

Interpret the results ➤ Values of $\bar{x}$ at least as great as 18.4 will be observed only about 0.82% of the time when a random sample of size 36 is taken from a population with mean 18 and standard deviation 1. The value $\bar{x} = 18.4$ is enough greater than 18 that the manufacturer's claim is questionable.

Other Cases

You now know a great deal about the sampling distribution of $\bar{x}$ in two cases: for a normal population distribution and for a large sample size. What happens when the population distribution is not normal and n is small? Although it is still true that $\mu_{\bar{x}} = \mu$ and $\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$, unfortunately there is no general result about the shape of the distribution.

When the objective is to draw a conclusion about the mean of such a population, one way to proceed is to make an assumption about the shape of the population distribution. Statisticians have proposed and studied a number of such models. Theoretical methods or simulation can be used to describe the sampling distribution of $\bar{x}$ corresponding to the assumed model.

An alternative strategy is to use one of the transformations presented in Chapter 7 to create a data set that more closely resembles a sample from a normal population and then to base inferences on the transformed data. Yet another strategy is to use procedures based on a statistic other than $\bar{x}$. Consult a statistician or a more advanced text for more information.

Exercises 8.10–8.22

8.10 A random sample is selected from a population with mean $\mu = 100$ and standard deviation $\sigma = 10$. Determine the mean and standard deviation of the sampling distribution of $\bar{x}$ for each of the following sample sizes:

a. $n = 9$

b. $n = 15$

c. $n = 36$

d. $n = 50$

e. $n = 100$

f. $n = 400$

8.11 For which of the sample sizes given in the previous exercise would it be reasonable to think that the sampling distribution of $\bar{x}$ is approximately normal in shape?

8.12 Explain the difference between σ and $\sigma_{\bar{x}}$ and between μ and $\mu_{\bar{x}}$.

8.13 The paper **"Privacy Policies for Apps Targeted Toward Youth: Descriptive Analysis of Readability" (*JIMR MHEALTH and UHEALTH* [2018]: 1–11)** describes a study of the readability of the privacy policies for apps that are targeted toward young people. The reading grade level (the grade level at which a person would be expected to be able to read and understand the text without help) was measured for the privacy policies of 30 free apps. The sample mean reading grade level was reported to be 13.0 (meaning that someone would need to be reading at grade level 13 [first year of college] to read and understand the policy). The sample standard deviation was 1.3. Suppose that it is reasonable to regard this sample of 30 free apps as a random sample of free apps targeted toward young people and that you are interested in learning about the value of μ, the population mean reading grade level for free apps targeted toward young people.

a. What is the approximate standard deviation of the sampling distribution of $\bar{x}$?

b. Is the sampling distribution of $\bar{x}$ approximately normal?

8.14 The paper referenced in the previous exercise also looked at the reading grade level for 34 apps that were available for purchase. The sample mean reading grade level was reported to be 12.5 (meaning that someone would need to be reading at grade level 12.5 [senior in high school] to read and understand the policy). The sample standard deviation was 1.8. Suppose that it is reasonable to regard this sample of 34 apps that are available for purchase as a random sample of apps available for purchase that are targeted toward young people and that you are interested in learning about the value of μ, the population mean reading grade level for apps available for purchase that are targeted toward young people.

a. What is the approximate standard deviation of the sampling distribution of $\bar{x}$?

Copyright 2025 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s). Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

b. Notice that the standard deviation of the sampling distribution of $\bar{x}$ is smaller for the free apps (see Exercise 8.13) than for the apps that were available for sale, even though the sample size for the free apps was smaller. Why did this occur?

c. Is the sampling distribution of $\bar{x}$ approximately normal?

8.15 Suppose that a random sample of size 64 is to be selected from a population with mean 40 and standard deviation 5.

a. What are the mean and standard deviation of the sampling distribution of $\bar{x}$? Describe the shape of the sampling distribution of $\bar{x}$.

b. What is the approximate probability that $\bar{x}$ will be within 0.5 of the population mean μ? (Hint: See Examples 8.5 and 8.6.)

c. What is the approximate probability that $\bar{x}$ will differ from μ by more than 0.7?

8.16 The time that a randomly selected individual waits for an elevator in an office building has a uniform distribution over the interval from 0 to 1 minute. For this distribution $\mu = 0.5$ and $\sigma = 0.289$.

a. Let $\bar{x}$ be the sample mean waiting time for a random sample of 16 individuals. What are the mean and standard deviation of the sampling distribution of $\bar{x}$?

b. Answer Part (a) but for a random sample of 50 individuals.

c. Draw a picture of the approximate sampling distribution of $\bar{x}$ when $n = 50$.

8.17 Let x denote the time (in minutes) that it takes a fifth-grade student to read a certain passage. Suppose that the mean value and standard deviation of x are $\mu = 2$ minutes and $\sigma = 0.8$ minutes, respectively.

a. If $\bar{x}$ is the sample mean time for a random sample of $n = 9$ students, where is the sampling distribution of $\bar{x}$ centered, and how much does it spread out around the center (as described by its standard deviation)?

b. Repeat Part (a) for a sample of size of $n = 20$ and again for a sample of size $n = 100$. How do the centers and variability of the three $\bar{x}$ distributions compare to one another?

c. Which of the sample sizes in Part (b) would be most likely to result in an $\bar{x}$ value close to μ, and why?

8.18 In the library on a university campus, there is a sign in the elevator that indicates a limit of 16 persons. In addition, there is a weight limit of 2500 pounds. Assume that the average weight of students, faculty, and staff on campus is 150 pounds, that the standard deviation is 27 pounds, and that the distribution of weights of individuals on campus is approximately normal. Suppose a random sample of 16 persons from the campus will be selected.

a. What is the mean of the sampling distribution of $\bar{x}$?

b. What is the standard deviation of the sampling distribution of $\bar{x}$?

c. What mean weights for a sample of 16 people will result in the total weight exceeding the weight limit of 2500 pounds?

d. What is the probability that a random sample of 16 people will exceed the weight limit?

8.19 Suppose that a sample of size 100 is to be drawn from a population with standard deviation 10.

a. What is the probability that the sample mean will be within 1 of the value of μ?

b. For this example ($n = 100$, $\sigma = 10$), complete each of the following statements by computing the appropriate value:

  i. Approximately 95% of the time, $\bar{x}$ will be within _____ of μ.

  ii. Approximately 0.3% of the time, $\bar{x}$ will be farther than _____ from μ.

8.20 A manufacturing process is designed to produce bolts with a 0.5-inch diameter. Once each day, a random sample of 36 bolts is selected and the bolt diameters are recorded. If the resulting sample mean is less than 0.49 inches or greater than 0.51 inches, the process is shut down for adjustment. The standard deviation for diameter is 0.02 inches.

What is the probability that the manufacturing line will be shut down unnecessarily? (Hint: Find the probability of observing an $\bar{x}$ in the shutdown range when the actual process mean really is 0.5 inches.)

8.21 College students with checking accounts typically write relatively few checks in any given month, whereas adults who are not students typically write many more checks during a month. Suppose that 50% of a bank's accounts are held by students and that 50% are held by adults who are not students. Let x denote the number of checks written in a given month by a randomly selected bank customer.

a. Make a sketch of what the probability distribution of x might look like.

b. Suppose that the mean value of x is 22.0 and that the standard deviation is 16.5. A random sample of $n = 100$ customers is to be selected. Where is the sampling distribution of $\bar{x}$ centered, and what is the standard deviation of the sampling distribution of $\bar{x}$? Sketch a rough picture of the sampling distribution.

c. Referring to Part (b), what is the approximate probability that $\bar{x}$ is at most 20? at least 25?

8.22 An airplane with room for 100 passengers has a total baggage limit of 6000 pounds. Suppose that the total weight of the baggage checked by an individual passenger is a random variable x with a mean value of 50 pounds and a standard deviation of 20 pounds. If 100 passengers will board a flight, what is the approximate probability that the total weight of their baggage will exceed the limit? (Hint: With $n = 100$, the total weight exceeds the limit when the mean weight $\bar{x}$ exceeds 6000/100.)

Section 8.3 The Sampling Distribution of a Sample Proportion

The objective of many statistical investigations is to draw a conclusion about the proportion of individuals or objects in a population that possess a specified property. For example, cell phones that don't require repair during the warranty period or coffee drinkers who regularly drink decaffeinated coffee. Because proportions are of interest in a variety of situations, it is helpful to introduce some general terminology. An individual or object that possesses the property of interest is labeled a success (S), and one that does not possess the property is termed a failure (F).

The letter p is used to denote the proportion of successes in the population. The value of p is a number between 0 and 1, and $100p$ is the percentage of successes in the population. For example, if $p = 0.75$, 75% of the population members are successes, and if $p = 0.01$, the population contains only 1% successes and 99% failures.

The value of p is usually unknown. When a random sample of size n is selected from a population that consists of successes and failures, some of the individuals in the sample are successes, and the rest are failures. The statistic that is used to draw conclusions about p is $\hat{p}$, the **sample proportion of successes**:

$$\hat{p} = \frac{\text{number of successes in the sample}}{n}$$

For example, if there are three successes in a sample of size 5, then $\hat{p} = 3/5 = 0.6$.

Drawing conclusions about p requires first learning about properties of the sampling distribution of the statistic $\hat{p}$. As was done for the sample mean, some simulation experiments will be carried out to develop an intuitive understanding of the sampling distribution of the sample proportion before stating general properties. In each example, 500 random samples of size n are selected from a population having a specified value of p. Then the value of $\hat{p}$ for each sample is calculated and a density histogram of the 500 $\hat{p}$ values is constructed.

Example 8.7 STEM College Students

In the fall of 2020, there were 21,447 students enrolled at California Polytechnic State University, San Luis Obispo. Of these students, 9439 (44%) were enrolled in a science, technology, engineering, or mathematics (STEM) major. To illustrate properties of the sampling distribution of a sample proportion, consider simulating taking samples from this population.

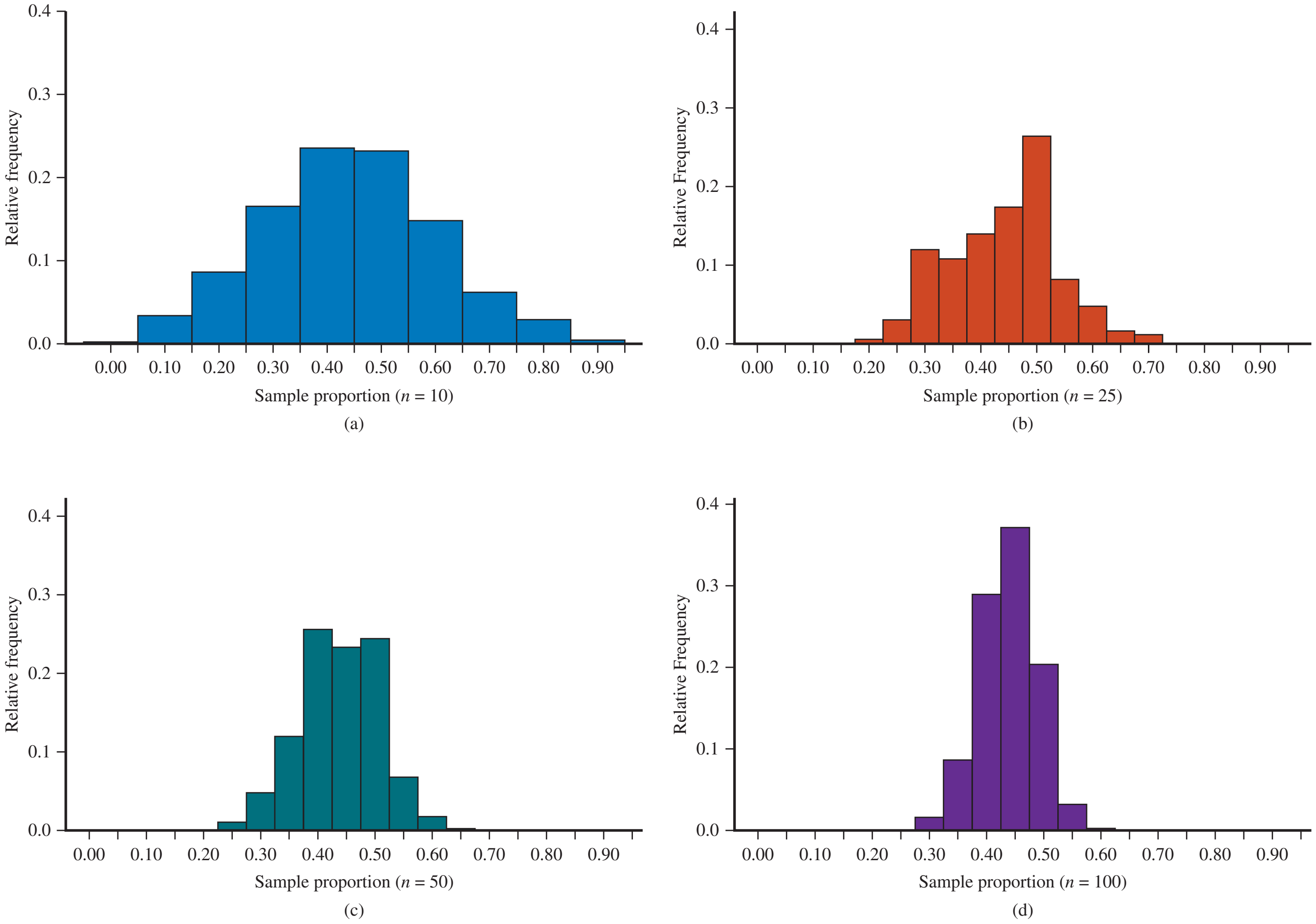
With success denoting a STEM major, the proportion of successes in the population is $p = 0.44$. A statistical software package was used to select 500 samples of size $n = 10$, then 500 samples of size $n = 25$, then 500 samples with $n = 50$, and finally 500 samples with $n = 100$. Density histograms of the 500 values of $\hat{p}$ for each of the four sample sizes are displayed in Figure 8.7.

The most noticeable feature of the histogram shapes is that all are approximately symmetric. All four histograms appear to be centered at roughly 0.44, the value of p for the population sampled. Had the histograms been based on all possible samples, each histogram would have been centered at exactly 0.44. Finally, as was the case with the sampling distribution of $\bar{x}$, the histograms spread out more for small sample sizes than for large sample sizes. Not surprisingly, the value of $\hat{p}$ based on a large sample size tends to be closer to p, the population proportion of successes, than does $\hat{p}$ from a small sample.

Copyright 2025 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s). Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

Figure 8.7

Density histograms for 500 values of $\hat{p}$ ($p = 0.44$) for Example 8.7: (a) $n = 10$; (b) $n = 25$; (c) $n = 50$; (d) $n = 100$.



(a) (b) (c) (d)

Example 8.8 Contracting Hepatitis from Blood Transfusion

The development of viral hepatitis after a blood transfusion can cause serious complications for a patient. The article **"An Assessment of Hepatitis E Virus in U.S. Blood Donors and Recipients" (*Transfusion* [2013]: 2505–2511)** reported that in a sample of 342 blood donors age 18 to 45, about 7% tested positive for the Hepatitis E virus. Suppose that the actual proportion that test positive for Hepatitis E in the population of all blood donors in this age group is 0.07. You can simulate sampling from this population of blood donors. For this example, a blood donor who tests positive for Hepatitis E will be considered a success, so $p = 0.07$. Figure 8.8 displays relative frequency histograms of 500 values of $\hat{p}$ for the four sample sizes $n = 10$, 25, 50, and 100.

As was the case in Example 8.7, all four histograms are centered at approximately the value of p for the population being sampled. (The average of the $\hat{p}$ values for these simulations are 0.0690, 0.0677, 0.0707, and 0.0694.) If the histograms had been based on all possible samples, they would all have been centered at exactly $p = 0.07$.

Notice that the scales on the axes are not the same for the four histograms in Figure 8.8. This was done so that it would be easier to see the behavior of the sample proportion for each sample size. Taking the differences in scales into account, you can see that the sample-to-sample variability decreases as the sample size increases. For example, the distribution is much less spread out in the histogram for $n = 100$ than for $n = 25$. The larger the value of n, the closer the sample proportion $\hat{p}$ tends to be to the value of the population proportion p.

Copyright 2025 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s). Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

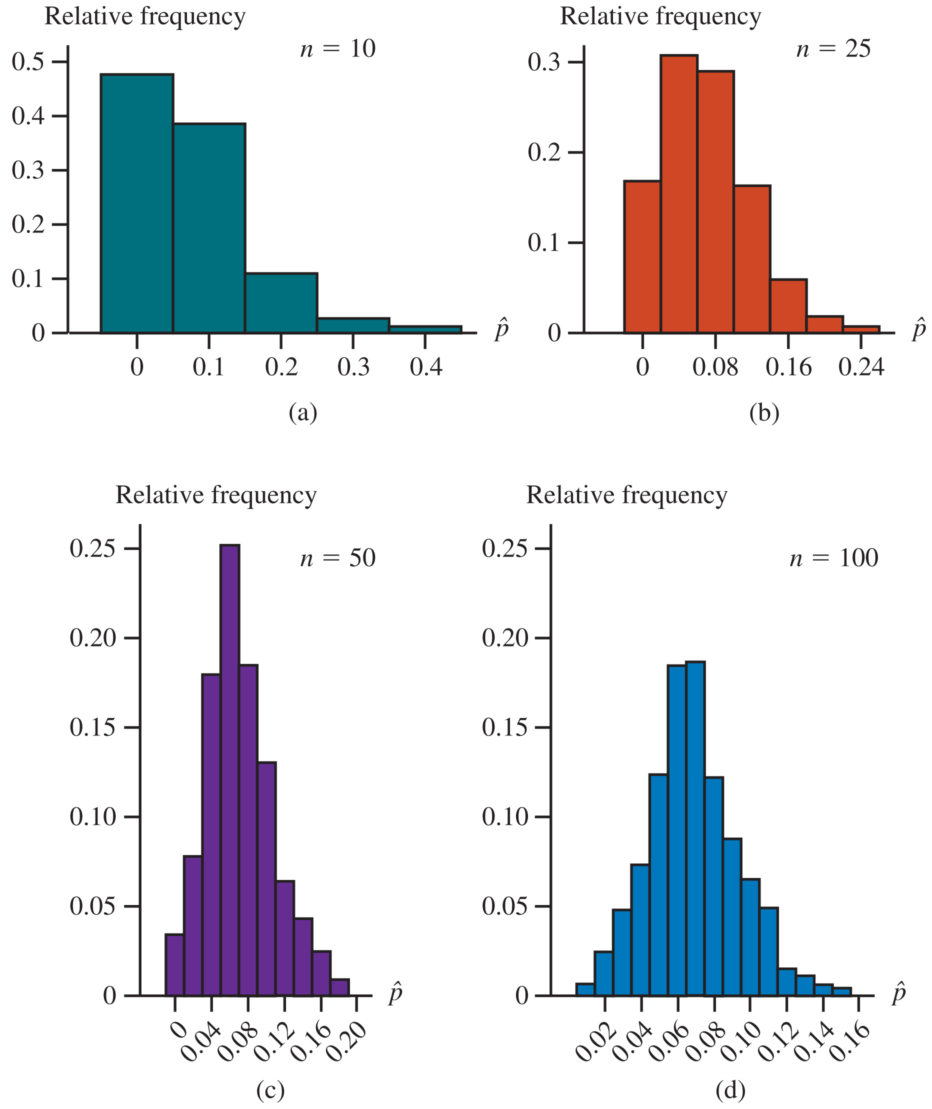


Figure 8.8
Histograms of 500 values of $\hat{p}$ ($p = 0.07$) for Example 8.8: (a) $n = 10$; (b) $n = 25$; (c) $n = 50$; (d) $n = 100$.

What Do You Think?

Which of the four sample sizes represented in Figure 8.8 resulted in a sampling distribution that is approximately normal?

Another thing to notice about the histograms in Figure 8.8 is the progression toward the shape of a normal curve as n increases. The histograms for $n = 10$ and $n = 25$ are quite skewed, and the skew of the histogram for $n = 50$ is still moderate (compare Figure 8.8(c) with Figure 8.7(c)). Only the histogram for $n = 100$ looks approximately normal in shape. It appears that whether a normal curve provides a good approximation to the sampling distribution of $\hat{p}$ depends on the values of both n and p. Knowing only that $n = 50$ is not enough to guarantee that the shape of the sampling distribution of $\hat{p}$ will be approximately normal.

General Properties of the Sampling Distribution of $\hat{P}$

Examples 8.7 and 8.8 suggest that the sampling distribution of $\hat{p}$ depends on both n, the sample size, and p, the proportion of successes in the population. Key results are stated more formally in the following general properties.

General Properties of the Sampling Distribution of $\hat{p}$

Notation

$\hat{p}$	Sample proportion
n	Sample size
p	Population proportion
$\mu_{\hat{p}}$	Mean of the sampling distribution of $\hat{p}$
$\sigma_{\hat{p}}$	Standard deviation of the sampling distribution of $\hat{p}$

When random samples are selected from a population, the following are properties of the sampling distribution of $\hat{p}$:

Property 1. $\mu_{\hat{p}} = p$.

Property 2. $\sigma_{\hat{p}} = \sqrt{\dfrac{p(1-p)}{n}}$.

This formula is exact if the population is infinite, and is approximately correct if the population is finite and no more than 10% of the population is included in the sample.

Property 3. When n is large and p is not too near 0 or 1, the sampling distribution of $\hat{p}$ is approximately normal.

The sampling distribution of $\hat{p}$ is always centered at the value of the population success proportion p, and the extent to which the distribution spreads out about p decreases as the sample size n increases.

Examples 8.7 and 8.8 indicate that both p and n must be considered in judging whether the sampling distribution of $\hat{p}$ is approximately normal.

The farther the value of p is from 0.5, the larger n must be in order for the sampling distribution of $\hat{p}$ to be approximately normal.

A conservative rule of thumb is that if both

$$\boldsymbol{np \geq 10 \text{ and } n(1-p) \geq 10},$$

then a normal distribution provides a reasonable approximation to the sampling distribution of $\hat{p}$.

A sample size of $n = 100$ is not by itself sufficient to justify the use of a normal approximation. If $p = 0.01$, the distribution of $\hat{p}$ is positively skewed even when $n = 100$, so a bell-shaped curve does not give a good approximation. Similarly, if $n = 100$ and $p = 0.99$ (so that $n(1-p) = 1 < 10$), the distribution of $\hat{p}$ has a substantial negative skew. The conditions $np \geq 10$ and $n(1-p) \geq 10$ ensure that the sampling distribution of $\hat{p}$ is not too skewed. If $p = 0.5$, the normal approximation can be used for n as small as 20, whereas for $p = 0.05$ or 0.95, n should be at least 200.

Example 8.9 Blood Transfusions Continued

Understand the context ➤

In the article referenced in Example 8.8, the proportion of blood donors testing positive for Hepatitis E was given as 0.07. Suppose that a new screening procedure is implemented and it is hoped that this will reduce the number of donors who are positive. Blood screened using this procedure is given to $n = 200$ blood recipients. Suppose that only 6 of the 200 patients contract hepatitis.

This appears to be a favorable result, because $\hat{p} = 6/200 = 0.03$. The question of interest to medical researchers is, Does this result indicate that the actual proportion of donors who are positive for Hepatitis E when the new screening procedure is used is less than 0.07, or could this result be plausibly attributed to sampling variability (that is, to the fact that $\hat{p}$ typically differs from the population proportion, p)?

Copyright 2025 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s). Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

If the screening procedure is not effective then p = 0.07, and

$$\mu_{\hat{p}} = 0.07$$

$$\sigma_{\hat{p}} = \sqrt{\frac{p(1-p)}{200}} = \sqrt{\frac{(0.07)(0.93)}{200}} = 0.018$$

Because

$$np = 200(0.07) = 14 \geq 10$$

and

$$n(1-p) = 200(0.93) = 186 \geq 10$$

the sampling distribution of $\hat{p}$ is approximately normal.

Do the work ➤ Then, if the screening procedure is not effective,

$$P(\hat{p} \leq 0.03) = P\left(z \leq \frac{0.03 - 0.07}{0.018}\right)$$
$$= P(z \leq -2.22)$$
$$= 0.0132$$

Interpret the results ➤ This small probability indicates that it is unlikely that a sample proportion 0.03 or smaller would be observed if the screening procedure was not effective. The new screening procedure appears to result in a smaller proportion of blood donors who are positive for Hepatitis E.

Exercises 8.23–8.31

8.23 A random sample is to be selected from a population that has a proportion of successes $p = 0.65$. Determine the mean and standard deviation of the sampling distribution of $\hat{p}$ for each of the following sample sizes:

a. $n = 10$
b. $n = 20$
c. $n = 30$
d. $n = 50$
e. $n = 100$
f. $n = 200$

8.24 **a.** For which of the sample sizes given in the previous exercise would the sampling distribution of $\hat{p}$ be approximately normal if $p = 0.65$?

b. For which of the sample sizes given in the previous exercise would the sampling distribution be approximately normal if $p = 0.2$?

8.25 Consider the two relative frequency histograms at the top of the next page. The histogram on the left was constructed by selecting 100 different random samples of size 35 from a population in which 17% donated to a nonprofit organization. For each sample, the sample proportion $\hat{p}$ was calculated and then the 100 $\hat{p}$ values were used to construct the histogram. The histogram on the right was constructed in a similar way but used samples of size 110.

a. Which of the two histograms indicates that the value of $\hat{p}$ has more sample-to-sample variability? How can you tell?

b. For which of the two sample sizes, $n = 35$ or $n = 110$, do you think the value of $\hat{p}$ would be more likely to be close to the actual population proportion of $p = 0.17$? What about the given histograms supports your choice?

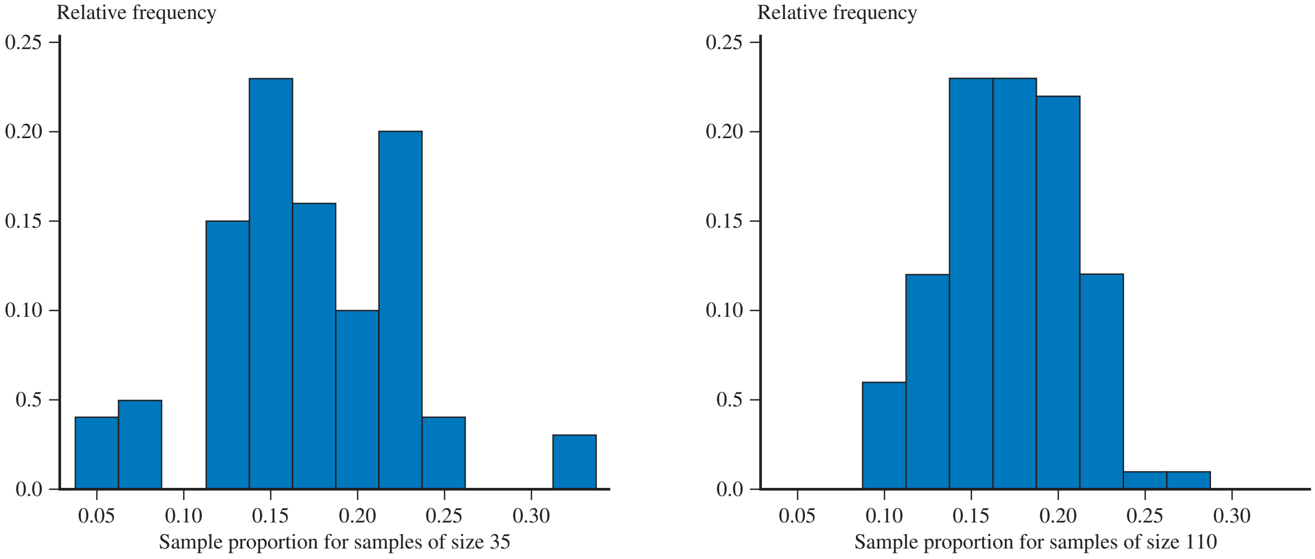
8.26 The article **"Keep It Clean: Social Media Screenings Gain in Popularity" (*Business News Daily*, November 1, 2021)** included data from a survey of 1012 hiring managers and human resource professionals. The article noted that many employers are using social networks to screen job applicants. Of the 1012 people who participated in the survey, 708 indicated that they use social networking sites to research job applicants. For the purposes of this exercise, assume that the sample can be regarded as a random sample of hiring managers and human resource professionals.

a. Suppose you are interested in learning about the value of p, the proportion of all hiring managers and human resource managers who use social networking sites to research job applicants. This proportion can be estimated using the sample proportion, $\hat{p}$. What is the value of $\hat{p}$ for this sample?

b. Based on what you know about the sampling distribution of $\hat{p}$, is it reasonable to think that this estimate is within 0.02 of the actual value of the population proportion? Explain why or why not.

Copyright 2025 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s). Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

Histograms for Exercise 8.25



8.27 The report **"A Crisis in Civic Education" (goacta.org, January 2016)** indicated that in a survey of a random sample of 1000 recent college graduates, 96 indicated that they believed that Judith Sheindlin (also known on TV as "Judge Judy") was a member of the U.S. Supreme Court. Is it reasonable to conclude that the proportion of recent college graduates who have this incorrect belief is greater than 0.09 (9%)?

8.28 The U.S. Census Bureau reported that in 2019, the proportion of adult Americans age 25 and older who have a bachelor's degree or higher is 0.36 **("Educational Attainment in the United States: 2019," census.gov)**. Consider the population of all adult Americans age 25 and over in 2015 and define $\hat{p}$ to the proportion of people in a random sample from this population who have a bachelor's degree or higher.

a. Would $\hat{p}$ based on a random sample of only 10 people from this population have approximately a normal distribution? Explain why or why not.

b. What are the mean value and standard deviation of $\hat{p}$ based on a random sample of size 400?

c. Suppose that the sample size is $n = 200$ rather than $n = 400$. Does the change in sample size affect the mean and standard deviation of the sampling distribution of $\hat{p}$? If so, what are the new values for the mean and standard deviation? If not, explain why not.

8.29 The article **"Fewer Americans Are Reading, But Don't Blame the Millennials" (*Los Angeles Times*, October 9, 2016)** indicates that 80% of millennials (those born between 1981 and 1996) have read a book in the last year. Suppose that this is the actual percentage for the population of all millennials. Consider a sample proportion $\hat{p}$ that is based on a random sample of 225 millennials.

a. If $p = 0.80$, what are the mean and standard deviation of the sampling distribution of $\hat{p}$?

b. If $p = 0.70$, what are the mean and standard deviation of the sampling distribution of $\hat{p}$?

c. Is the sampling distribution of $\hat{p}$ approximately normal in both cases ($p = 0.80$ and $p = 0.70$)?

8.30 Suppose that a candidate for public office is favored by 48% of all registered voters in his district. A polling organization will take a random sample of 500 voters and will use $\hat{p}$, the sample proportion, to estimate p. What is the approximate probability that $\hat{p}$ will be greater than 0.5, causing the polling organization to incorrectly predict the result of the upcoming election? (Hint: See Example 8.9.)

8.31 The article **"Nearly a Quarter of Americans Say Their Zoom Fatigue Is Higher Than Ever—Can Dogs Help?" (HonestPaws Blog, October 22, 2021)** summarizes results from a survey of 400 participants. For these respondents, 51% said that seeing a coworker's dog in a Zoom meeting makes the meeting more enjoyable for them. Suppose that this sample can be regarded as a random sample of working adults. Is it reasonable to conclude that more than half of all working adults feel this way? Use what you know about the sampling distribution of $\hat{p}$ to support your answer.

Copyright 2025 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s). Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

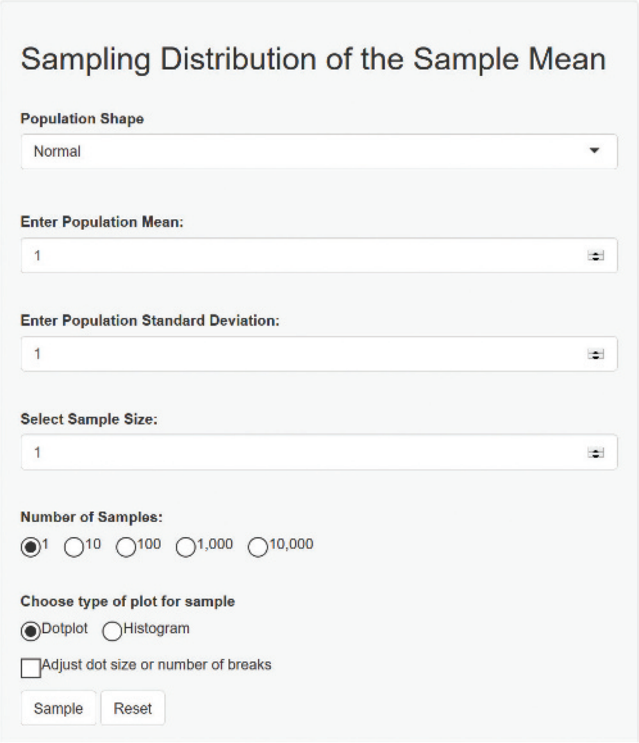
Chapter Activities

Activity 8.1 Sampling Distribution of the Sample Mean Technology Activity

The purpose of this activity is to use technology to explore properties of the sampling distribution of the sample mean.

This activity uses the Shiny app titled Sampling Distribution Mean that is part of the collection of Shiny web apps that accompany this text. This app can be found at www.cengage.com/Peck.

Open the app. You should see a screen like the one shown here.



Sampling Distribution of the Sample Mean

Population Shape

Normal

Enter Population Mean:

1

Enter Population Standard Deviation:

1

Select Sample Size:

1

Number of Samples:

● 1 ○ 10 ○ 100 ○ 1,000 ○ 10,000

Choose type of plot for sample

● Dotplot ○ Histogram

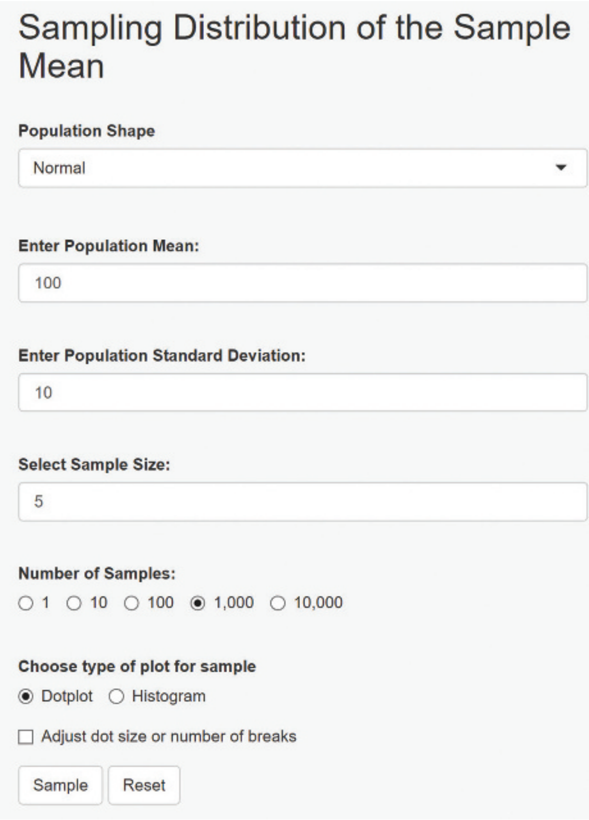
☐ Adjust dot size or number of breaks

Sample Reset

This app can be used to carry out sampling experiments like the ones described in Examples 8.2 and 8.3. In those examples, random samples were selected from a specified population to learn about the behavior of the sample mean $\bar{x}$. This app allows you to specify a population distribution shape, as well as some characteristics of the population distribution, such as the mean. You can also specify a sample size and the number of samples you would like to generate.

Part 1: Normal Population

Step 1. Start by looking at the distribution of sample means when random samples are selected from a population that has a normal distribution with mean $\mu = 100$ and standard deviation $\sigma = 10$. Make sure that “Normal” is selected as the population shape and enter 100 for the population mean and 10 for the population standard deviation. Begin by entering 5 for the sample size and select 1000 for the number of samples to generate. Choose dotplot for the type of display for the sample.



Sampling Distribution of the Sample Mean

Population Shape

Normal

Enter Population Mean:

100

Enter Population Standard Deviation:

10

Select Sample Size:

5

Number of Samples:

○ 1 ○ 10 ○ 100 ● 1,000 ○ 10,000

Choose type of plot for sample

● Dotplot ○ Histogram

☐ Adjust dot size or number of breaks

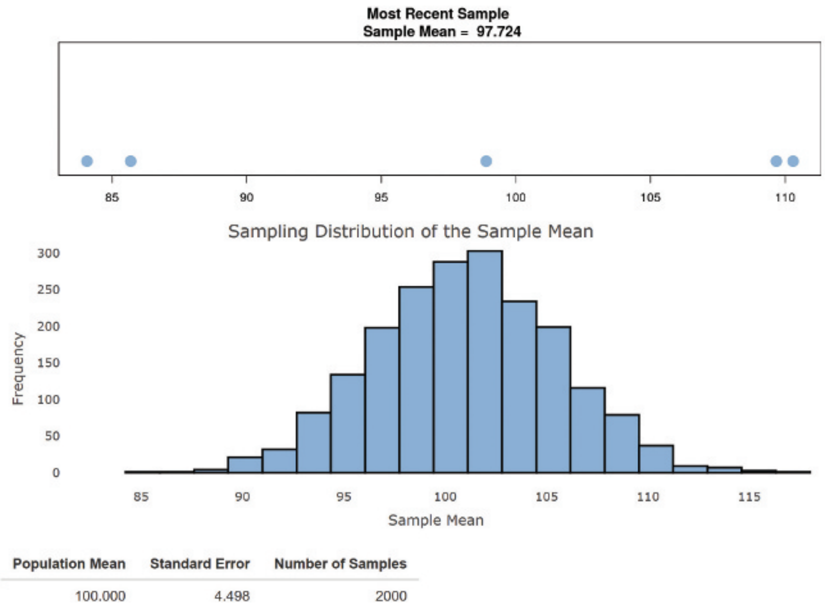
Sample Reset

When you click on the sample button, the app will generate 1000 random samples of size 5 from this normal population, calculate the sample mean for each sample, and display a histogram of the sample means. You should see a screen similar to the one that follows, although yours will be different since you will not have the same set of 1000 random samples.

Write a brief description of the histogram of the sample means, being sure to discuss center, variability, and shape.

Step 2. Next, click the reset button to clear the display. Generate a new histogram of 1000 sample means for random samples of size 10. Write a brief description of the histogram of the sample means, being sure to discuss center, variability, and shape. How does this distribution compare to the distribution of sample means from Step 1?

Step 3. Repeat Step 2 for sample sizes of 20 and 50.


Most Recent Sample
Sample Mean = 97.724
85
90
95
100
105
110
Sampling Distribution of the Sample Mean
Frequency
300
250
200
150
100
50
0
85
90
95
100
105
110
115
Sample Mean

Population Mean	Standard Error	Number of Samples
100.000	4.498	2000

Step 4. Consider the four distributions of sample means that you created (corresponding to sample sizes of 5, 10, 20, and 50). Explain how these histograms are consistent with the general properties of the sampling distribution of the sample mean given in Section 8.2.

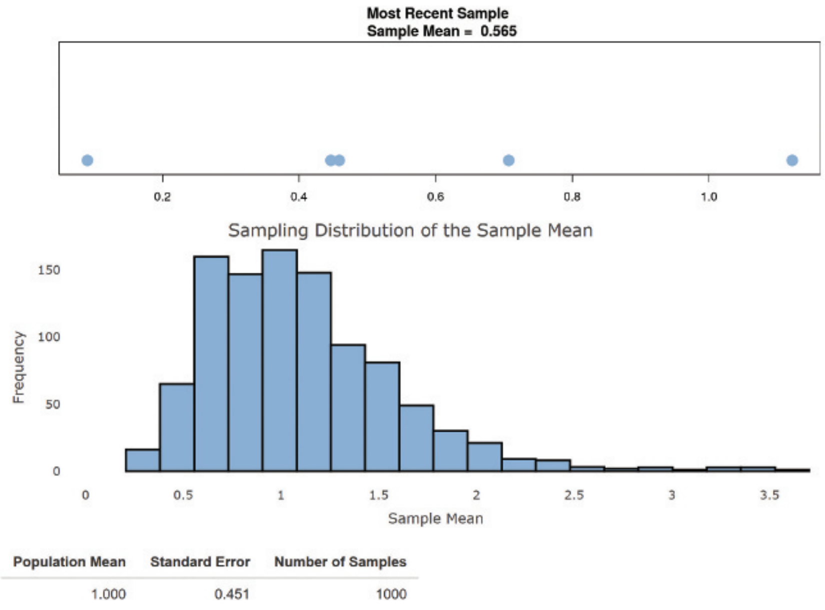
Part 2: A Skewed Population

Step 1. In Part 2, you will consider the distribution of sample means when random samples are selected from a population that is skewed. Begin by selecting "Exponential" as the population shape and enter 1 for the population mean. Enter 5 for the sample size and select 1000 for the number of samples to generate. Choose dotplot for the type of display for the sample. When you click on the sample button, the app will generate 1000 random samples of size 5 from this population, calculate the sample mean for each sample, and display a histogram of the sample means. You should see a screen similar to the one shown here, although yours will be different since you will not have the same set of 1000 random samples. Notice that the population distribution shown in the top graph is skewed.

Write a brief description of the histogram of the sample means, being sure to discuss center, variability, and shape.

Step 2. Next click the reset button to clear the display. Generate a new histogram of 1000 sample means for random samples of size 10. Write a brief description of the histogram of the sample means, being sure to discuss center, variability, and shape. How does this distribution compare to the distribution of sample means from Step 1?

Step 3. Repeat Step 2 for sample sizes of 20 and 50.


Most Recent Sample
Sample Mean = 0.565
0.2
0.4
0.6
0.8
1.0
Sampling Distribution of the Sample Mean
Frequency
150
100
50
0
0
0.5
1
1.5
2
2.5
3
3.5
Sample Mean

Population Mean	Standard Error	Number of Samples
1.000	0.451	1000

Copyright 2025 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s). Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

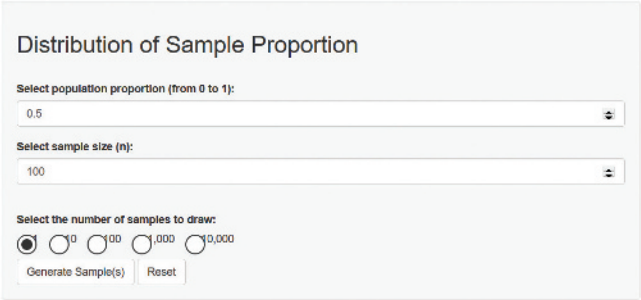
Step 4. Consider the four distributions of sample means that you created (corresponding to sample sizes of 5, 10, 20, and 50). Explain how these histograms are consistent with the general properties of the sampling distribution of the sample mean given in Section 8.2.

Activity 8.2 Sampling Distribution of the Sample Proportion Technology Activity

The purpose of this activity is to use technology to explore properties of the sampling distribution of the sample proportion.

This activity uses the Shiny app titled Sampling Distribution Proportion that is part of the collection of Shiny web apps that accompany this text. This app can be found at **www.cengage.com/Peck**.

In this activity, you will explore properties of the sampling distribution of the sample proportion. Open the app. You should see a screen like the one shown here.



Distribution of Sample Proportion

Select population proportion (from 0 to 1): 0.5

Select sample size (n): 100

Select the number of samples to draw:

● 1 ○ 10 ○ 100 ○ 1,000 ○ 10,000

This app can be used to carry out sampling experiments like the ones described in Examples 8.7 and 8.8. In those examples, random samples were selected from a specified population to learn about the behavior of the sample proportion $\hat{p}$. This app allows you to specify a population proportion. You can also specify a sample size and the number of samples you would like to generate.

Part 1: Exploring the effect of sample size

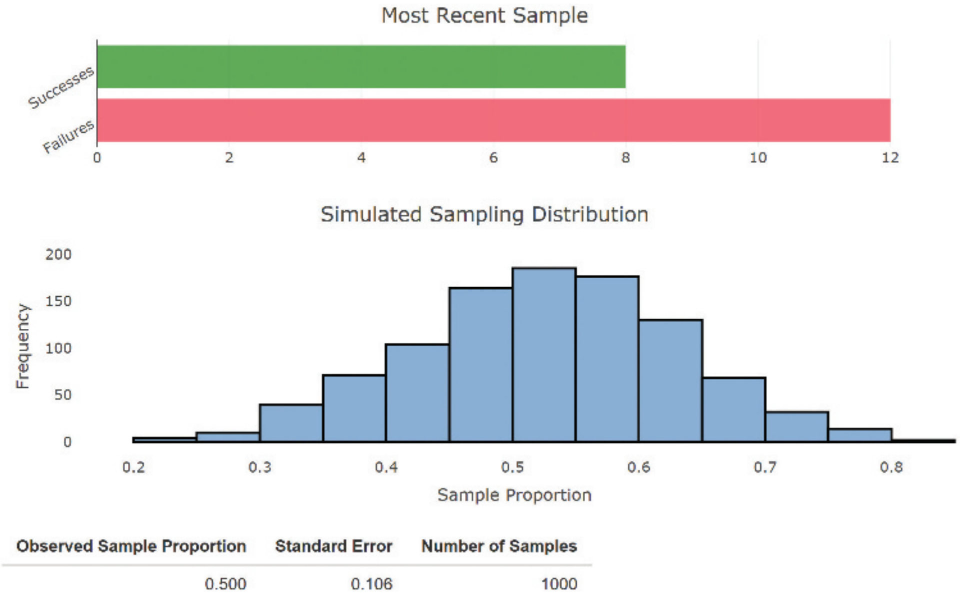
Step 1. Start by looking at the distribution of sample proportions when random samples are selected from a population that has a proportion of successes of 0.5. Enter 0.5 for the population proportion. Begin by entering 20 for the sample size and select 1000 for the number of samples to generate. When you click on the Generate Samples button, the app will generate 1000 random samples of size 20 from this population, calculate the sample proportion for each sample, and display a histogram of the sample proportions. You should see a screen similar to the one shown here, although yours will be different since you will not have the same set of 1000 random samples.

Write a brief description of the histogram of the sample proportions, being sure to discuss center, variability, and shape.

Step 2. Next, click the reset button to clear the display. Generate a new histogram of 1000 sample proportions for random samples of size 30. Write a brief description of the histogram of the sample proportions, being sure to discuss center, variability, and shape. How does this distribution compare to the distribution of sample proportions from Step 1?

Step 3. Repeat Step 2 for sample sizes of 50 and 100.

Step 4. Consider the four distributions of sample proportions that you created (corresponding to sample sizes of 20, 30, 50, and 100). Explain how these histograms are consistent with the general properties of the sampling distribution of the sample proportion given in Section 8.3.



Observed Sample Proportion	Standard Error	Number of Samples
0.500	0.106	1000

Copyright 2025 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s). Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

Part 2: Exploring the effect of value of the population proportion

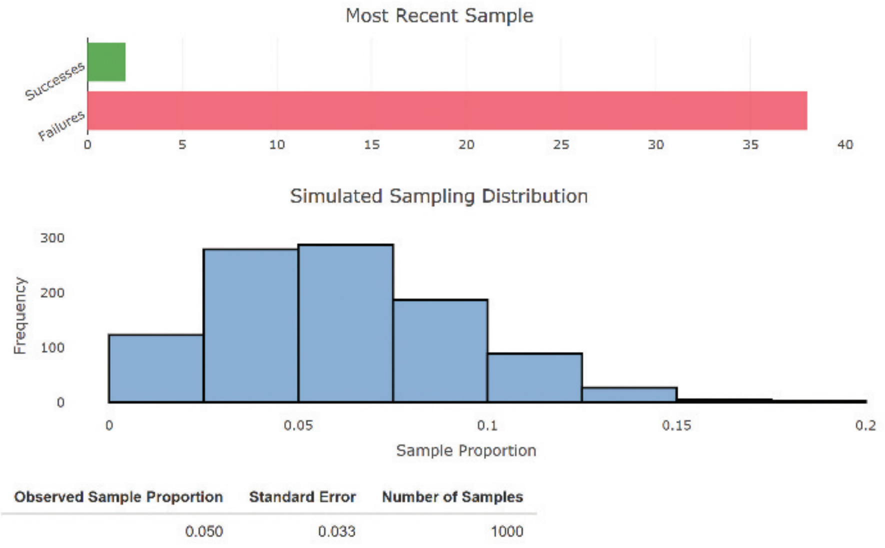
Step 1. Start by looking at the distribution of sample proportions when random samples of size 40 are selected from a population that has a proportion of success of 0.05. Enter 0.05 for the population proportion and 40 for the sample size. Select 1000 for the number of samples to generate. When you click on the Generate Samples button, the app will generate 1000 random samples of size 40 from this population, calculate the sample proportion for each sample, and display a histogram of the sample proportions. You should see a screen similar to the one shown here, although yours will be different since you will not have the same set of 1000 random samples.

Write a brief description of the histogram of the sample proportions, being sure to discuss center, variability, and shape.

Step 2. Next, click the reset button to clear the display. Generate a new histogram of 1000 sample proportions for random samples of size 40 from a population with a proportion of successes of 0.3. Write a brief description of the histogram of the sample proportions, being sure to discuss center, variability, and shape. How does this distribution compare to the distribution of sample proportions from Step 1?

Step 3. Repeat Step 2 for population proportions of 0.5 and 0.8.

Step 4. Consider the four distributions of sample proportions that you created (corresponding to population proportions of 0.05, 0.3, 0.5, and 0.8). Explain how these histograms are consistent with the general properties of the sampling distribution of the sample proportion given in Section 8.3.



Observed Sample Proportion	Standard Error	Number of Samples
0.050	0.033	1000

Activity 8.3 Do Students Who Take the SATs Multiple Times Have an Advantage in College Admissions? Technology Activity

The purpose of this activity is to use what you have learned about sampling distributions to evaluate different admission policies and make a recommendation.

This activity requires use of a computer or a graphing calculator.

Background: ***The Chronicle of Higher Education* (January 29, 2003)** summarized an article that appeared on the ***American Prospect*** website titled **"College Try: Why Universities Should Stop Encouraging Applicants to Take the SATs Over and Over Again."** This paper argued that current college admission policies that permit applicants to take the SAT exam multiple times and then use the highest score for consideration of admission favor students from families with higher incomes (who can afford to take the exam many times). The author proposed two alternatives that he believes would be fairer than using the highest score: (1) Use the average of all test scores, or (2) use only the most recent score.

In this activity, you will investigate the differences between the three possibilities by looking at the sampling distributions of three statistics for a test taker who takes the exam twice and for a test taker who takes the exam five times. The three statistics are

Max = maximum score
Mean = average score
Recent = most recent score

An individual's score on the SAT exam fluctuates between test administrations. Suppose that a particular student's "true ability" is reflected by an SAT score of 1200 but, because of chance fluctuations, the test score on any particular administration of the exam can be considered a random variable that has a distribution that is approximately normal with mean 1200 and standard deviation 30. If you select a sample from this normal distribution, the resulting

Copyright 2025 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s). Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

set of observations can be viewed as a collection of test scores that might have been obtained by this student.

Part 1: Begin by considering what happens if this student takes the exam twice. You will use simulation to generate samples of two test scores, Score1 and Score2, for this student. Then you will compute the values of Max, Mean, and Recent for each pair of scores. The resulting values of Max, Mean, and Recent will be used to construct approximations to the sampling distributions of the three statistics.

The instructions that follow assume the use of Minitab. If you are using a different software package or a graphing calculator, your instructor will provide alternative instructions.

a. Obtain 500 sets of two test scores by generating observations from a normal distribution with mean 1200 and standard deviation 30.

Minitab: Calc → Random Data → Normal
- Enter 500 in the Generate box (to get 500 sets of scores)
- Enter C1-C2 in the Store in Columns box (to get two test scores in each set)
- Enter 1200 in the Mean box (because we want scores from a normal distribution with mean 1200)
- Enter 30 in the Standard Deviation box (because we want scores from a normal distribution with standard deviation 30)
- Click on OK

b. Looking at the Minitab worksheet, you should now see 500 rows of values in each of the first two columns. The two values in any particular row can be regarded as the test scores that might be observed when the student takes the test twice. For each pair of test scores, you can now calculate the values of Max, Mean, and Recent.

i. Recent is just the last test score, so the values in C2 are the values of Recent. Name this column recent2 by typing the name into the gray box at the top of C2.

ii. Compute the maximum test score (Max) for each pair of scores, and store the values in C3, as follows:

Minitab: Calc → Row statistics
- Click the button for maximum
- Enter C1-C2 in the Input variables box
- Enter C3 in the Store Result In box.
- Click on OK

You should now see the maximum value for each pair in C3. Name this column max2.

iii. Compute the average test score (Mean) for each pair of scores, and store the values in C4, as follows:

Minitab: Calc → Row statistics
- Click the button for mean
- Enter C1-C2 in the Input Variables box
- Enter C4 in the Store Result In box.
- Click on OK

You should now see the average for each pair in C4. Name this column mean2.

c. Construct density histograms for each of the three statistics (these density histograms approximate the sampling distributions of the three statistics), as follows:

Minitab: Graph → Histogram
- Enter max2, mean2, and recent2 into the first three rows of the Graph Variables box
- Click on the Options button. Select Density. Click on OK. (This will produce histograms that use the density scale rather than the frequency scale.)
- Click on the Frame drop-down menu, and select Multiple Graphs. Select Same X and Same Y. (This will cause Minitab to use the same scales for all three histograms, so that they can be easily compared.)
- Click on OK.

Part 2: Now you will produce approximate sampling distributions for these same three statistics, but for the case of a student who takes the exam five times. Follow the same steps as in Part 1, with the following modifications:

a. Obtain 500 sets of five test scores, and store these values in columns C11– C15.

b. Recent will just be the values in C15; name this column recent5. Compute the Max and Mean values, and store them in columns C16 and C17. Name these columns max5 and mean5.

c. Construct density histograms for max5, mean5, and recent5.

Part 3: Now use the approximate sampling distributions constructed in Parts 1 and 2 to answer the following questions.

a. The statistic that is the average of the test scores is just a sample mean (for a sample of size 2 in Part 1 and for a sample of size 5 in Part 2). How do the sampling distributions of mean2 and mean5 compare to what is expected based on the general properties of the $\bar{x}$ distribution given in Section 8.2? Explain.

b. Based on the three distributions from Part 1, for a two-time test taker, describe the advantage of using the maximum score compared to using either the average score or the most recent score.

c. Now consider the approximate sampling distributions of the maximum score for two-time and for five-time test takers. How do these two distributions compare?

d. Does a student who takes the exam five times have a big advantage over a student of equal ability who takes the exam only twice if the maximum score is used for college admission decisions? Explain.

e. If you were writing admission procedures for a selective university, would you recommend using the maximum test score, the average test score, or the most recent test score in making admission decisions? Write a paragraph explaining your choice.

Copyright 2025 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s). Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

Summary Key Concepts and Formulas

Term or Formula	Comment
Statistic	Any quantity whose value is calculated from sample data.
Sampling distribution	The probability distribution of a statistic. The sampling distribution describes the long-run behavior of the statistic.
Sampling distribution of $\bar{x}$	The probability distribution of the sample mean $\bar{x}$ based on a random sample of size n. Properties of the $\bar{x}$ sampling distribution: $\mu_{\bar{x}} = \mu$ and $\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$ (where μ and σ are the population mean and standard deviation, respectively). In addition, when the population distribution is normal or the sample size is large, the sampling distribution of $\bar{x}$ is (approximately) normal.
Central Limit Theorem	This important theorem states that when n is sufficiently large, the $\bar{x}$ distribution will be approximately normal. The standard rule of thumb is that the theorem can safely be applied when n is greater than or equal to 30.
Sampling distribution of $\hat{p}$	The probability distribution of the sample proportion $\hat{p}$, based on a random sample of size n. When the sample size is sufficiently large, the sampling distribution of $\hat{p}$ is approximately normal, with $\mu_{\hat{p}} = p$ and $\sigma_{\hat{p}} = \sqrt{\frac{p(1-p)}{n}}$ where p is the value of the population proportion.

Chapter Review Exercises 8.32–8.37

8.32 Consider the two relative frequency histograms at the bottom of the page. The histogram on the left was constructed by selecting 100 different random samples of size 40 from a population consisting of 20% part-time students and 80% full-time students. For each sample, the sample proportion of part-time students, $\hat{p}$, was calculated. The 100 $\hat{p}$ values were used to construct the histogram. The histogram on the right was constructed in a similar way, but using samples of size 70.

a. Which of the two histograms indicates that the value of $\hat{p}$ has smaller sample-to-sample variability? How can you tell?

b. For which of the two sample sizes, $n = 40$ or $n = 70$, do you think the value of $\hat{p}$ would be *less* likely to be close to 0.20? What about the given histograms supports your choice?

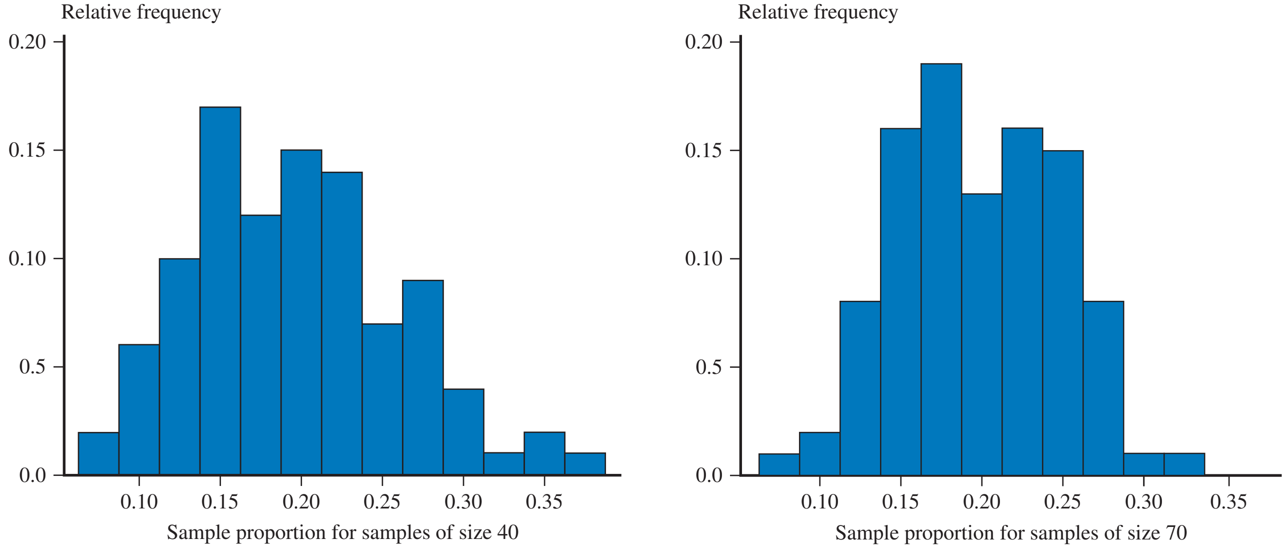
8.33 Let $x_1, x_2, \ldots, x_{100}$ denote the actual weights (in pounds) of 100 randomly selected bags of fertilizer. Suppose that the weight of a randomly selected bag has a distribution with mean 50 pounds and variance 1 pound^2. Let $\bar{x}$ be the sample mean weight ($n = 100$).

a. Describe the sampling distribution of $\bar{x}$.

b. What is the probability that the sample mean is between 49.75 and 50.25 pounds?

c. What is the probability that the sample mean is less than 50 pounds?

Histograms for Exercise 8.32



8.34 Based on a survey of 7267 adults, the National Retail Federation estimated that 27% of adults will buy a Valentine's Day gift for their pet **("Confident Consumers and Broader Buying Lead to Record Valentine's Day Spending Plans," National Retail Federation, January 30, 2020)**. Suppose that it is reasonable to regard this sample as a random sample of adults. Is it reasonable to conclude that the proportion of all adults who will buy a Valentine's Day gift for their pet is greater than 0.25? Use what you know about the sampling distribution of $\hat{p}$ to support your answer.

8.35 The article **"More Americans Have Tattoos Today Than Seven Years Ago" (Ipsos, August 29, 2019)** describes a survey of 1005 adult Americans. Suppose that it is reasonable to regard this sample as a random sample from the population of adult Americans. Of the survey participants, 295 reported that they had at least one tattoo.

a. Suppose you are interested in learning about the value of p, the proportion of all adult Americans who have at least one tattoo. This proportion can be estimated using the sample proportion, $\hat{p}$. What is the value of $\hat{p}$ for this sample?

b. Based on what you know about the sampling distribution of $\hat{p}$, is it reasonable to think that this estimate is within 0.05 of the actual value of the population proportion? Explain why or why not.

8.36 Suppose that 40% of all U.S. employees contribute to a retirement plan ($p = 0.40$).

a. In a random sample of 100 employees, what is the approximate probability that at least half of those in the sample contribute to a retirement plan?

b. Suppose you were told that at least 60 of the 100 employees in a sample from your state contribute to a retirement plan. Would you think $p = 0.40$ for your state? Explain.

8.37 The amount of money spent by a customer at a discount store has a mean of \$100 and a standard deviation of \$30. What is the probability that a randomly selected group of 50 shoppers will spend a total of more than \$5300? (Hint: The total will be more than \$5300 when the sample mean exceeds what value?)

Chapter 9

Estimation Using a Single Sample

EF Stock/Shutterstock.com

College students use cell phones in multiple ways. The paper **"Sleeping with Technology: Cognitive, Affective and Technology Usage Predictors of Sleep Problems Among College Students" (*Sleep Health* [2016]: 49–56)** describes the results of a survey of a representative sample of 734 college students. Of the 734 students surveyed, 125 reported that they sleep with their cell phones near their bed and check their phones for something other than the time at least twice during the night. The authors of this paper were interested in using data from the sample to learn about the proportion of all college students who check their cell phones for something other than the time at least twice during the night.

The methods introduced in this chapter can be used to estimate this proportion. Because the estimate will be based only on a sample rather than on a census of all college students, it is important that this estimate be constructed in a way that also conveys information about the anticipated accuracy.

Learning Objectives

After completing this chapter, you will be able to:

- Explain what it means for a statistic to be an unbiased estimator of a population characteristic.
- Describe the relationships between sample size, margin of error, and the width of a confidence interval.
- Explain the meaning of the confidence level associated with a confidence interval.
- Construct and interpret a confidence interval for a population proportion.
- Describe the factors that affect the width of a confidence interval.
- Construct and interpret a confidence interval for a population mean.
- Determine the sample size necessary to achieve a desired margin of error when estimating a population proportion or a population mean.
- Calculate and interpret a bootstrap confidence interval for a population proportion. (Optional)
- Calculate and interpret a bootstrap confidence interval for a population mean. (Optional)

Copyright 2025 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s). Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

Section 9.1 Point Estimation

A simple way to estimate a population characteristic involves using sample data to calculate a single number that represents a plausible value of the population characteristic. For example, sample data might suggest that 1000 hours is a plausible value for μ, the mean lifetime for all lightbulbs of a particular brand. A survey of a sample of students at a particular university might lead to the statement that 0.41 is a plausible value for p, the proportion of all students at the university who favor a fee increase to expand recreational facilities.

Definition

Point estimate: A single number calculated using sample data that represents a plausible value of a population characteristic.

A point estimate is obtained by first selecting an appropriate statistic. The value of the statistic for a given sample is then used as the point estimate. For example, the calculated value of the sample mean is one point estimate of a population mean μ, and the sample proportion is a point estimate of a population proportion, p.

Example 9.1 Sleeping with Your Cell Phone

Understand the context ➤

● One of the purposes of the survey described in the chapter introduction was to estimate the proportion of college students who check their phone for something other than the time more than twice during the night. Based on information given in the paper, 125 of the 734 students surveyed reported checking their phone more than twice during the night. You can use this information to estimate p, where p is the proportion of all college students who check their phone more than twice during the night.

With a success identified as a student who checks their phone more than twice during the night, p is the population proportion of successes. The statistic

$$\hat{p} = \frac{\text{number of successes in the sample}}{n}$$

Do the work ➤

which is the sample proportion of successes, is an obvious choice for obtaining a point estimate of p. Using the reported information, the point estimate of p is

$$\hat{p} = \frac{125}{734} = 0.170$$

Interpret the results ➤

Based on this random sample, you would estimate that 17.0% of college students check their phone more than twice during the night.

For purposes of estimating a population proportion p, there is no obvious alternative to the statistic $\hat{p}$. In other situations, such as the one illustrated in Example 9.2, there may be several statistics that might be used to obtain an estimate.

Example 9.2 Electric Car Energy Consumption

Understand the context ➤

The Environmental Protection Agency introduced “miles per gallon-equivalent” (MPG-e) as a way to compare the amount of energy consumed by electric and other alternative fuel vehicles to gas-powered vehicles. MPG-e is a measure of how far an electric vehicle can travel on 33.7 kWh of electricity, which is the approximately equivalent to one gallon of gas. This means that a rating of 40 MPG-e for an electric car is comparable to a gas-powered car that gets 40 miles per gallon of gas.

Consider the data ➤

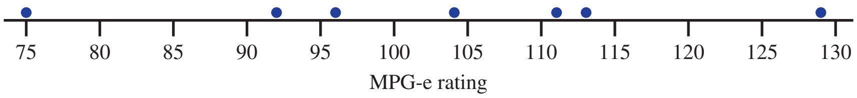
The U.S. Department of Energy reports MPG-e for electric vehicles sold in the United States (**fueleconomy.gov**). The accompanying data are the MPG-e ratings for a random sample of seven 2021 model electric cars.

75 92 96 113 129 111 104

● **Data set available online**

Copyright 2025 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s). Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

A dotplot of the data is shown here:



From the dotplot, you can see that the distribution of MPG-e rating is approximately symmetric.

Suppose a point estimate of μ, the mean MPG-e rating for all 2021 electric vehicle models, is desired. An obvious choice of a statistic for estimating μ is the sample mean, $\bar{x}$. However, there are other possibilities. If the population distribution is symmetric, the population mean and the population median are equal. Since the dotplot of the sample is approximately symmetric, you might consider using the sample median as an estimate of the population mean.

The estimates of μ based on the mean and the median for this sample are

Do the work ➤

$$\text{sample mean} = \bar{x} = \frac{\Sigma x}{n} = \frac{720}{7} = 102.86$$

$$\text{sample median} = 104$$

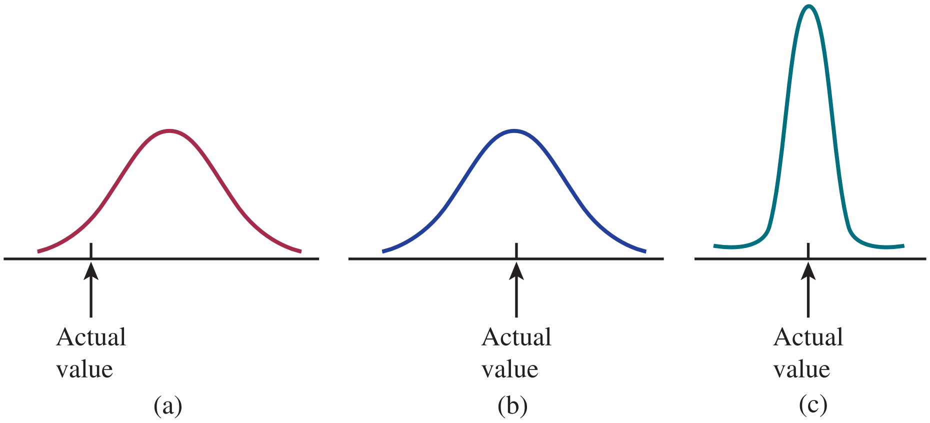
These two estimates of the mean MPG-e rating for 2021 electric car models are different. The choice between them should depend on which statistic tends, on average, to produce an estimate closest to the actual value of μ. The following subsection provides some criteria for choosing among competing statistics.

Choosing a Statistic for Calculating an Estimate

As illustrated in Example 9.2, there may be more than one statistic that is a reasonable choice as a point estimate of a population characteristic. You would like to use a statistic that tends to produce an accurate estimate—that is, an estimate close to the actual value of the population characteristic. Information about the accuracy of estimation for a particular statistic is provided by the statistic's sampling distribution.

Figure 9.1 displays the sampling distributions of three different statistics. The value of the population characteristic, which is denoted by *actual value* in the figure, is marked on the measurement axis. The distribution in Figure 9.1(a) is that of a statistic unlikely to result in an estimate close to the actual value. The distribution is centered to the right of the actual value, making it very likely that an estimate (a value of the statistic for a particular sample) will be larger than the actual value. If this statistic is used to calculate an estimate based on a first sample, then another estimate based on a second sample, and another estimate based on a third sample, and so on, the long-run average value of these estimates will be greater than the actual value.

Figure 9.1
Sampling distributions of three different statistics for estimating a population characteristic.



(a) (b) (c)

Copyright 2025 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s). Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

The sampling distribution of Figure 9.1(b) is centered at the actual value. This means that although one estimate may be smaller than the actual value and another may be larger, when this statistic is used many times over with different random samples, there will be no long-run tendency to overestimate or underestimate the actual value. Notice that even though the sampling distribution is correctly centered, it spreads out quite a bit around the actual value. Because of this, some estimates resulting from the use of this statistic will be far above or far below the actual value, even though there is no systematic tendency to underestimate or overestimate the actual value.

In contrast, the mean value of the statistic with the distribution shown in Figure 9.1(c) is equal to the actual value of the population characteristic (implying no systematic tendency to overestimate or underestimate the actual value). The statistic's standard deviation is relatively small. This means that estimates based on this third statistic will almost always be quite close to the actual value—certainly more often than estimates resulting from the statistic with the sampling distribution shown in Figure 9.1(b).

Definitions

Unbiased statistic: A statistic whose mean value is equal to the value of the population characteristic being estimated.

Biased statistic: A statistic that is not unbiased.

What Do You Think?

Which of the three sampling distributions pictured in Figure 9.1 are for statistics that are unbiased?

As an example of a statistic that is biased, consider using the sample range as an estimate of the population range. Because the range of a population is defined as the difference between the largest value in the population and the smallest value, the range for a sample tends to underestimate the population range. This is because the largest value in a sample must be less than or equal to the largest value in the population and the smallest sample value must be greater than or equal to the smallest value in the population. The value of the sample range equals the value of the population range *only* if the sample happens to include both the largest and the smallest values in the population. In all other cases, the value of the sample range is smaller than the population range. This means that $\mu_{\text{sample range}}$ is less than the value of the population range, and the sample range is a biased statistic for estimating the population range.

One of the general results concerning the sampling distribution of $\bar{x}$, the sample mean, is that $\mu_{\bar{x}} = \mu$. This result says that the $\bar{x}$ values from all possible random samples of size n center around μ, the population mean. For example, if $\mu = 100$, the $\bar{x}$ distribution is centered at 100, whereas if $\mu = 5200$, then the $\bar{x}$ distribution is centered at 5200. This means that $\bar{x}$ is an unbiased statistic for estimating μ. Similarly, because the sampling distribution of $\hat{p}$ is centered at p, it follows that $\hat{p}$ is an unbiased statistic for estimating a population proportion.

Using an unbiased statistic that also has a small standard deviation ensures that there will be no systematic tendency to underestimate or overestimate the value of the population characteristic *and* that estimates will almost always be relatively close to the actual value of the population characteristic.

If there are several unbiased statistics that might be used for estimating a population characteristic, the best choice among them is the statistic with the smallest standard deviation.

What Do You Think?

Can you think of a situation where you might prefer a biased statistic over an unbiased statistic? (Hint: think about the standard deviation of the estimator.)

Consider the problem of estimating a population mean, μ. An obvious choice of a statistic for obtaining a point estimate of μ is the sample mean, $\bar{x}$, an unbiased statistic for estimating the population mean. However, when the population distribution is symmetric, $\bar{x}$ is not the only unbiased statistic for estimating the population mean. For example, another unbiased statistics for estimating μ in this case is the sample median. Which statistic should be used? If the population distribution is normal, then $\bar{x}$ has a smaller standard deviation than any other unbiased statistic for estimating μ. This means that if the population distribution is normal, the sample mean would be the best choice.

Now consider estimating another population characteristic, the population variance, σ^2. The sample variance

$$s^2 = \frac{\Sigma(x - \bar{x})^2}{n - 1}$$

is a good choice for obtaining a point estimate of the population variance, σ^2. It can be shown that s^2 is an unbiased statistic for estimating σ^2. This means that whatever the value of σ^2, the sampling distribution of s^2 is centered at that value. This is the reason that the divisor $(n - 1)$ is used. An alternative statistic is the average squared deviation

$$\frac{\Sigma(x - \bar{x})^2}{n}$$

which has a more natural denominator than s^2. However, the average squared deviation is biased, with its values tending to be smaller, on average, than the actual value of σ^2.

Example 9.3 Airborne Times for Flights from San Francisco to Washington, D.C.

Understand the context ➤

● The **Bureau of Transportation Statistics** provides data on U.S. airline flights. Suppose that the airborne times (in minutes) for nonstop flights from San Francisco to Washington Dulles airport for 10 randomly selected flights in June 2023 are:

270 256 267 285 274 275 266 258 271 281

For these data $\Sigma x = 2703$, $\Sigma x^2 = 731{,}373$, $n = 10$, and

$$\begin{aligned}\Sigma(x - \bar{x})^2 &= \Sigma x^2 - \frac{(\Sigma x)^2}{n}\\ &= 731{,}373 - \frac{(2703)^2}{10}\\ &= 752.1\end{aligned}$$

Do the work ➤

σ^2 denotes the actual variance of airborne time for June 2023 nonstop flights from San Francisco to Washington Dulles airport. Using the sample variance s^2 to provide a point estimate of σ^2 yields

$$s^2 = \frac{\Sigma(x - \bar{x})^2}{n - 1} = \frac{752.1}{9} = 83.57$$

Using the average squared deviation (with divisor $n = 10$), the resulting point estimate is

$$\frac{\Sigma(x - \bar{x})^2}{n} = \frac{752.1}{10} = 75.21$$

Interpret the results ➤

Because s^2 is an unbiased statistic for estimating σ^2, many statisticians would recommend using the point estimate 83.57.

An obvious choice of a statistic for estimating the population standard deviation σ is the sample standard deviation s. For the data given in Example 9.3,

$$s = \sqrt{83.57} = 9.14$$

● **Data set available online**

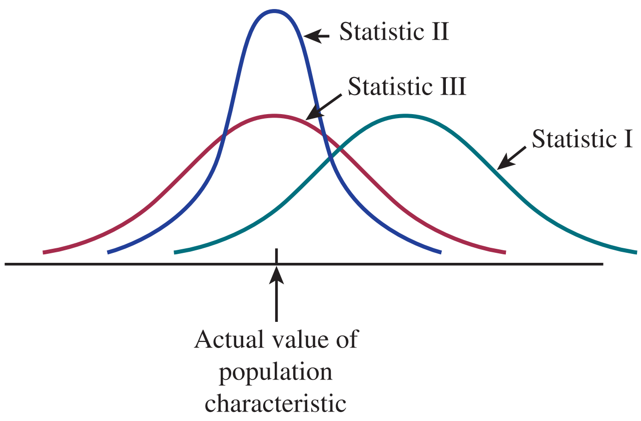
Copyright 2025 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s). Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

Unfortunately, the fact that s^2 is an unbiased statistic for estimating σ^2 does not mean that s is an unbiased statistic for estimating σ. The sample standard deviation tends to underestimate slightly the actual value of σ. However, unbiasedness is not the only criterion by which a statistic can be judged, and there are other good reasons for using s to estimate σ. For this reason, whenever an estimate of σ based on a single random sample is needed, the the statistic s will be used to obtain a point estimate.

Exercises 9.1–9.9

● **Data set available online**

9.1 Three different statistics are being considered for estimating a population characteristic. The sampling distributions of the three statistics are shown in the following illustration:



Which statistic would you recommend? Explain your choice. (Hint: See the section on choosing a statistic.)

9.2 **a.** Why is an unbiased statistic generally preferred over a biased statistic for estimating a population characteristic?

b. Does unbiasedness alone guarantee that the estimate will be close to the true value? Explain.

c. Under what circumstances might you choose a biased statistic over an unbiased statistic if two statistics are available for estimating a population characteristic?

9.3 The report **"The 2016 Consumer Financial Literacy Survey" (The National Foundation for Credit Counseling (nfcc.org)** summarized data from a representative sample of 1668 adult Americans. When asked if they typically carry credit card debt from month to month, 584 of these people responded "yes." Estimate p, the proportion of adult Americans who carry credit card debt from month to month.

9.4 ● The authors of the paper **"Influence of Biofeedback Weight Bearing Training in Sit to Stand to Sit and the Limits of Stability on Stroke Patients" (*The Journal of Physical Therapy Science* [2016]: 3011–2014)** randomly selected two samples of patients admitted to the hospital after suffering a stroke. One sample was selected from patients who received biofeedback weight training for 8 weeks and the other sample was selected from patients who did not receive this training. At the end of 8 weeks, the time it took (in seconds) to stand from a sitting position and then to sit down again (called sit-stand-sit time) was measured for the people in each sample. Data consistent with summary quantities given in the paper are given below. For purposes of this exercise, you can assume that the samples are representative of the population of stroke patients who receive the biofeedback training and the population of stroke patients who do not receive this training.

Biofeedback Group

1.9	2.6	4.3	2.1	2.7	4.1	3.2	4.0
3.2	3.5	2.8	3.5	3.5	2.3	3.1	

No Biofeedback Group

5.1	4.7	3.9	4.2	4.7	4.3	4.2	5.1
3.4	4.2	5.1	4.4	4.0	3.4	3.9	

a. Use the given data to estimate the mean sit-stand-sit time for stroke patients who receive biofeedback training.

b. Use the given data to estimate the mean sit-stand-sit time for stroke patients who do not receive biofeedback training.

c. Use the given information to estimate the standard deviation of sit-stand-sit times for stroke patients who receive biofeedback.

9.5 Each person in a random sample of 20 students at a particular university was asked whether they are registered to vote. The responses (R = registered, N = not registered) are given here:

R R N R N N R R R N R R R R R N R R R N

Use these data to estimate p, the proportion of all students at the university who are registered to vote.

9.6 Suppose that each of 935 smokers received a nicotine patch, which delivers nicotine to the bloodstream but at a much slower rate than cigarettes.

Dosage was decreased to 0 over a 12-week period. Suppose that 245 of the subjects were still not smoking 6 months after treatment. Assuming it is reasonable to regard this sample as representative of all smokers, estimate the proportion of all smokers who, when given this treatment, would refrain from smoking for at least 6 months.

9.7 ● Given below are the sodium contents (in mg) for seven brands of hot dogs rated as "very good" by ***Consumer Reports*** **(consumerreports.org)**:

420 470 350 360 270 550 530

a. Use the given data to produce a point estimate of μ, the true mean sodium content for hot dogs.
b. Use the given data to produce a point estimate of σ^2, the variance of sodium content for hot dogs.
c. Use the given data to produce an estimate of σ, the standard deviation of sodium content. Is the statistic you used to produce your estimate unbiased? (Hint: See the discussion following Example 9.3.)

9.8 ● Suppose that a random sample of $n = 12$ four-year-old red pine trees was selected, and the diameter (in inches) of each tree's main stem was measured. The resulting observations are given here:

11.3 10.7 12.4 15.2 10.1 12.1 16.2 10.5
11.4 11.0 10.7 12.0

a. Calculate a point estimate of σ, the population standard deviation of main stem diameter. What statistic did you use to obtain your estimate?
b. Suppose that the diameter distribution is normal. Then the 90th percentile of the diameter distribution is $\mu + 1.28\sigma$ (so 90% of all trees have diameters less than this value). Calculate a point estimate for this percentile. (Hint: First calculate an estimate of μ and then use it along with your estimate of σ from Part (a).)

9.9 ● Suppose that a random sample of 10 houses heated with natural gas in a particular area is selected, and the amount of gas (in therms) used during the month of January is determined for each house. The resulting observations are as follows:

103 156 118 89 125 147 122 109 138 99

a. Let μ_J denote the average gas usage during January by all houses in this area. Calculate a point estimate of μ_J.
b. Suppose that 10,000 houses in this area use natural gas for heating. Let τ denote the total amount of gas used by all of these houses during January. Estimate τ using the given data. What statistic did you use to calculate your estimate?
c. Use the given data to estimate p, the proportion of all houses that used at least 100 therms.
d. Give a point estimate of the population median usage based on the given data. Which statistic did you use?

Section 9.2 Large-Sample Confidence Interval for a Population Proportion

In Section 9.1, you saw how to use a statistic to produce a point estimate of a population characteristic. The value of a point estimate depends on which sample, out of all the possible samples, happens to be selected. Different samples usually produce different estimates as a result of chance differences from one sample to another. Because of sampling variability, the point estimate from a sample is not usually exactly equal to the actual value of the population characteristic. However, you would hope that the chosen statistic produces an estimate that is close, on average, to the true value.

Although a point estimate may represent your best single-number estimate for the value of the population characteristic, it is not the only plausible value. As an alternative to a point estimate, the sample data can be used to report an interval of plausible values for the population characteristic. For example, you might be confident that for all text messages sent from cell phones, the proportion p of messages that are longer than 50 characters is in the interval from 0.53 to 0.57. The narrowness of this interval implies that you have rather precise information about the value of p. If, with the same confidence, you could only state that p was between 0.32 and 0.74, you don't really know much about the value of p.

Definition

Confidence interval: An interval of plausible values for a population characteristic.

A confidence interval is constructed so that you have a chosen level of confidence that the actual value of the population characteristic will be between the lower endpoint and the upper endpoint of the interval.

Associated with each confidence interval is a **confidence level**. The confidence level provides information on how much "confidence" you can have in the *method* used to construct the interval estimate (*not* your confidence in any one particular interval).

Usual choices for confidence levels are 90%, 95%, and 99%, although other confidence levels are also possible. If you were to construct a 95% confidence interval using the method to be described shortly, you would be using a method that is "successful" 95% of the time. This means that if this method was used to generate an interval estimate over and over again with different random samples, in the long run 95% of the resulting intervals would include the actual value of the characteristic being estimated. Similarly, a 99% confidence interval is one that is constructed using a method that is, in the long run, successful in capturing the actual value of the population characteristic 99% of the time.

Definition

Confidence level: The success rate of the *method* used to construct a confidence interval.

Many statistical studies are carried out to estimate the proportion of individuals or objects in a population that possess a particular property of interest (whatever has been defined as a "success"). For example, a university administrator might be interested in the proportion of students who prefer a new registration system to the previous registration method. In a different setting, a quality control engineer might be interested in the proportion of defective parts produced by a particular manufacturing process.

Recall that p denotes the proportion of individuals in the population that possess the property of interest (successes). Previously, the sample proportion

$$\hat{p} = \frac{\text{number in the sample that possess the property of interest}}{n}$$

was used to calculate a point estimate of p. You can also use $\hat{p}$ to construct a confidence interval for p.

Although a small-sample confidence interval for p can be obtained, in this text, only the large-sample case is considered. The construction of the large-sample interval is based on properties of the sampling distribution of the statistic $\hat{p}$:

1. The sampling distribution of $\hat{p}$ is centered at p. This means that $\mu_{\hat{p}} = p$, so $\hat{p}$ is an unbiased statistic for estimating p.
2. As long as the sample size is less than 10% of the population size, the standard deviation of $\hat{p}$ is well approximated by $\sigma_{\hat{p}} = \sqrt{\frac{p(1-p)}{n}}$
3. As long as n is large ($np \geq 10$ and $n(1-p) \geq 10$), the sampling distribution of $\hat{p}$ is well approximated by a normal curve.

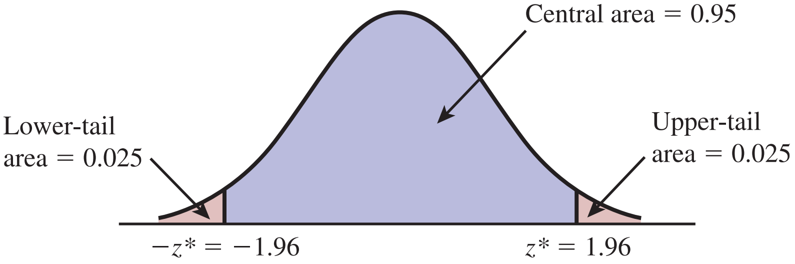
The accompanying box summarizes these properties.

When n is large and the sample size is less than 10% of the population size, the statistic $\hat{p}$ has a sampling distribution that is approximately normal with mean p and standard deviation $\sqrt{\frac{p(1-p)}{n}}$.

Copyright 2025 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s). Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

The development of a confidence interval for p is easier to follow when a particular confidence level is considered first. For a confidence level of 95%, technology or the table of standard normal (z) curve areas (Appendix Table 2) can be used to determine a value z^* such that a central area of 0.95 falls between $-z^*$ and z^*. In this case, the remaining area of 0.05 is divided equally between the two tails, as shown in Figure 9.2.

Figure 9.2
Capturing a central area of 0.95 under the z curve.



Generalizing this result to normal distributions other than the standard normal distribution tells you that for *any* normal distribution, about 95% of the values are within 1.96 standard deviations of the mean. For large random samples, the sampling distribution of $\hat{p}$ is approximately normal with mean $\mu_{\hat{p}} = p$ and standard deviation $\sigma_{\hat{p}} = \sqrt{\frac{p(1-p)}{n}}$, which leads to the following result.

When n is large, approximately 95% of all random samples of size n will result in a value of $\hat{p}$ that is within $1.96\,\sigma_{\hat{p}} = 1.96\sqrt{\frac{p(1-p)}{n}}$ of the actual value of the population proportion p.

If $\hat{p}$ is within $1.96\sqrt{\frac{p(1-p)}{n}}$ of p, this means the interval

$$\hat{p} - 1.96\sqrt{\frac{p(1-p)}{n}} \quad \text{to} \quad \hat{p} + 1.96\sqrt{\frac{p(1-p)}{n}}$$

will capture p. This will happen for 95% of all possible samples. However, if $\hat{p}$ is farther away from p than $1.96\sqrt{\frac{p(1-p)}{n}}$ (which will happen for about 5% of all possible samples), the interval will not include the value of p. This is shown in Figure 9.3.

Because $\hat{p}$ is within $1.96\sigma_{\hat{p}}$ of p for 95% of all possible random samples, this means that in repeated random sampling, 95% of the intervals

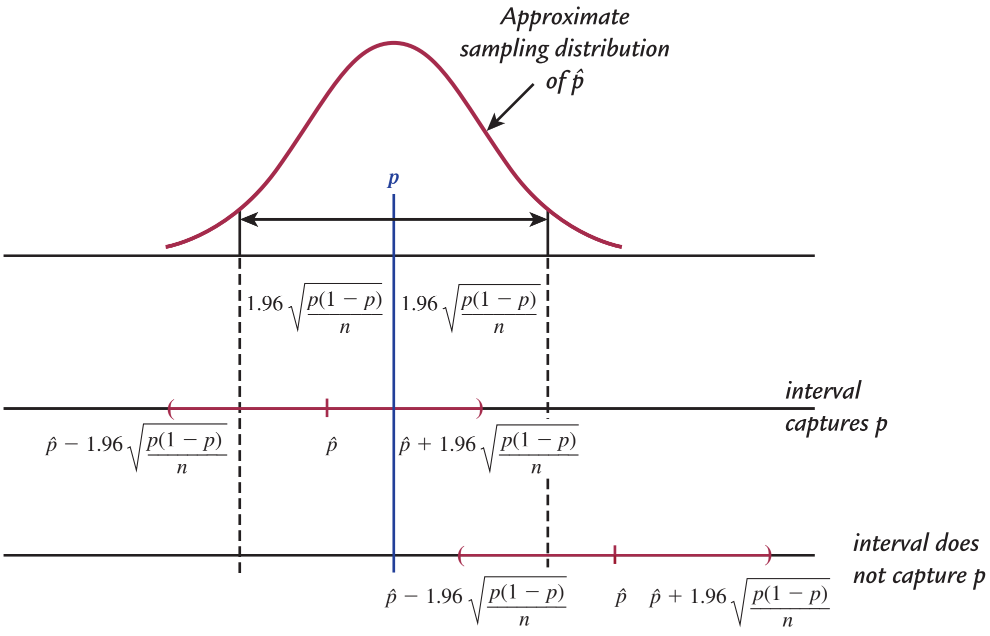
$$\hat{p} - 1.96\sqrt{\frac{p(1-p)}{n}} \quad \text{to} \quad \hat{p} + 1.96\sqrt{\frac{p(1-p)}{n}}$$

will contain the value of p.

Since the value of p is unknown, $\sqrt{\frac{p(1-p)}{n}}$ must be estimated. As long as the sample size is large, the value of $\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$ can be used in place of $\sqrt{\frac{p(1-p)}{n}}$.

Copyright 2025 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s). Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

Figure 9.3
The population proportion p is captured in the interval from $\hat{p} - 1.96\sqrt{\frac{p(1-p)}{n}}$ to $\hat{p} + 1.96\sqrt{\frac{p(1-p)}{n}}$ when $\hat{p}$ is within $1.96\sqrt{\frac{p(1-p)}{n}}$ of p.


Approximate sampling distribution of $\hat{p}$
p
$1.96\sqrt{\frac{p(1-p)}{n}}$
$1.96\sqrt{\frac{p(1-p)}{n}}$
interval captures p
$\hat{p} - 1.96\sqrt{\frac{p(1-p)}{n}}$
$\hat{p}$
$\hat{p} + 1.96\sqrt{\frac{p(1-p)}{n}}$
interval does not capture p
$\hat{p} - 1.96\sqrt{\frac{p(1-p)}{n}}$
$\hat{p}$
$\hat{p} + 1.96\sqrt{\frac{p(1-p)}{n}}$

When n is large, a **95% confidence interval for p** is

$$\left(\hat{p} - 1.96\sqrt{\frac{\hat{p}(1-\hat{p})}{n}},\ \hat{p} + 1.96\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}\right)$$

This can also be written as

$$\hat{p} \pm 1.96\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

where $\hat{p} + 1.96\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$ gives the upper endpoint of the interval and $\hat{p} - 1.96\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$ gives the lower endpoint of the interval.

This interval can be used as long as

1. $n\hat{p} \geq 10$ and $n(1-\hat{p}) \geq 10$,
2. the sample size is less than 10% of the population size if sampling is without replacement, and
3. it is reasonable to regard the sample as a random sample from the population of interest.

Example 9.4 Shark Bites

Understand the context ➤

Sharks are an important part of the marine ecosystem, but many people are afraid of sharks. The authors of the paper **"Influence of Social Media on Fear of Sharks Perceptions of Intentionality Associated with Shark Bites, and Shark Management Preferences" (*Frontiers in Communication*, October 2022)** studied efforts to reduce fear of sharks and change public misconceptions. As part of their study, they asked a sample of 340 adult residents of North Carolina who they thought was to blame when sharks bite people. Possible responses were sharks, swimmers, no one, the government, and other. Of those surveyed, 39 responded that they thought sharks were to blame. For purposes of this example, assume that it is reasonable to regard this sample as a representative sample of adult residents of North Carolina.

Copyright 2025 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s). Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

With p denoting the proportion of all adult residents of North Carolina who believe that sharks are to blame when sharks bite people, a point estimate of p is

$$\hat{p} = \frac{39}{340} = 0.115$$

Before calculating a 95% confidence interval to estimate p, you should check to make sure that the three necessary conditions are met:

1. $n\hat{p} = 340(0.115) = 39$ and $n(1 - \hat{p}) = 340(1 - 0.115) = 340(0.885) = 301$ are both greater than or equal to 10, so the sample size is large enough to proceed.
2. The sample size of $n = 340$ is much smaller than 10% of the population size (the number of adult residents of North Carolina).
3. For this example, it is assumed that it is reasonable to regard the sample as a random sample from the population.

Formulate a plan ➤ Because all three conditions are met, it is appropriate to use the sample data to construct a 95% confidence interval for p.

Do the work ➤ A 95% confidence interval for p is

$$\begin{aligned}\hat{p} \pm 1.96\sqrt{\frac{\hat{p}(1-\hat{p})}{n}} &= 0.115 \pm 1.96\sqrt{\frac{(0.115)(1-0.115)}{340}} \\ &= 0.115 \pm (1.96)(0.017) \\ &= 0.115 \pm 0.033 \\ &= (0.082, 0.148)\end{aligned}$$

Interpret the results ➤ Based on this sample, you can be 95% confident that p, the proportion of adult residents of North Carolina who believe sharks are to blame when sharks bite people, is between 0.082 and 0.148. A *method* was used to construct this estimate that will successfully capture the actual value of p 95% of the time in the long run.

The 95% confidence interval for p calculated in Example 9.4 is (0.082, 0.148). It is tempting to say that there is a "probability" of 0.95 that p is between 0.082 and 0.148. *Do not yield to this temptation!* The 95% refers to the percentage of *all* possible samples resulting in an interval that includes the value of p. In other words, if you take random sample after random sample from the population and use each one separately to calculate a 95% confidence interval, in the long run roughly 95% of these intervals will capture the value of p.

Figure 9.4 illustrates this concept for intervals generated from 100 different random samples. In this particular set of 100 intervals, 93 include p, whereas 7 do not. Any specific interval, and our interval (0.082, 0.148) in particular, either includes p or it does not (remember, the value of p is a fixed number but not known). You cannot make a chance (probability) statement concerning this particular interval. The confidence level 95% refers to the method used to construct the interval rather than to any particular interval, such as the one obtained in Example 9.4.

The formula given for a 95% confidence interval can easily be adapted for other confidence levels. The choice of a 95% confidence level led to the use of the z value 1.96 (chosen to capture a central area of 0.95 under the standard normal curve) in the confidence interval formula. Any other confidence level can be obtained by using an appropriate z critical value in place of 1.96.

For example, suppose that you wanted to achieve a confidence level of 99%. To obtain a central area of 0.99, the appropriate z critical value (using technology or Appendix Table 2) is $z = 2.58$, as illustrated in Figure 9.5. A 99% confidence interval for p is then obtained by using 2.58 in place of 1.96 in the formula for the 95% confidence interval.

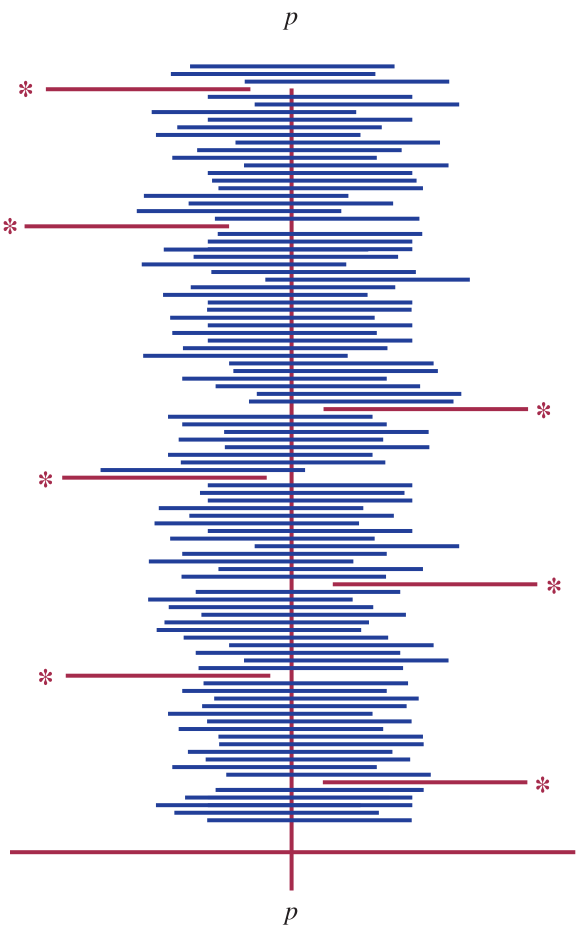


Figure 9.4
One hundred 95% confidence intervals for p calculated from 100 different random samples (asterisks identify intervals that do not include p).

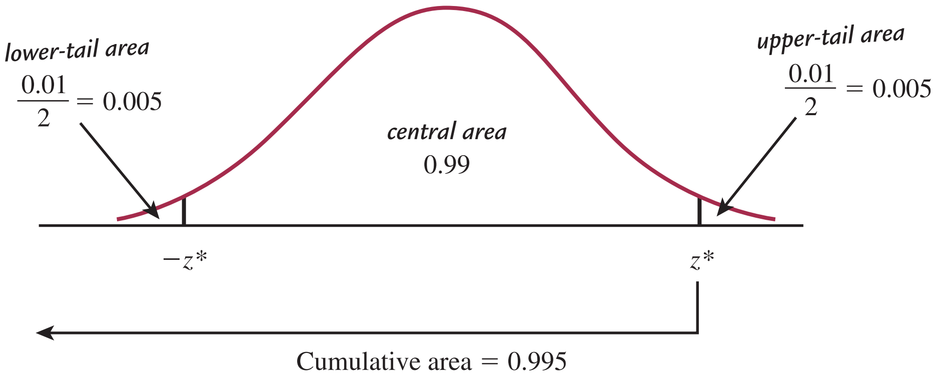


Figure 9.5
Finding the z critical value for a 99% confidence level.

What Do You Think?

Using the same sample data, which would be wider—a 95% confidence interval or a 99% confidence interval?

Why settle for 95% confidence when 99% confidence is possible? Because the higher confidence level comes at a price. The resulting interval is wider than the 95% interval. The width of the 95% interval is $2\left(1.96\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}\right)$, whereas the 99% interval has width $2\left(2.58\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}\right)$. The higher *reliability* of the 99% interval (where "reliability" is specified by the confidence level) entails a loss in precision (as indicated by the wider interval). In the opinion of many investigators, a 95% confidence interval produces a reasonable compromise between reliability and precision.

Copyright 2025 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s). Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

The Large-Sample Confidence Interval for *p*

The general formula for a confidence interval for a population proportion p when

1. $\hat{p}$ is the sample proportion from a **simple random sample**,
2. the sample size ***n* is large** ($n\hat{p} \geq 10$ and $n(1-\hat{p}) \geq 10$), and
3. if the sample is selected without replacement, **the sample size is small relative to the population size** (n is at most 10% of the population size)*

is

$$\hat{p} \pm (z \text{ critical value})\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

The desired confidence level determines which z critical value is used. The three most commonly used confidence levels, 90%, 95%, and 99%, use z critical values 1.645, 1.96, and 2.58, respectively.

Note: This interval is not appropriate for small samples. It is possible to construct a confidence interval in the small-sample case. For one method, see optional Section 9.5.

*In Chapter 7, you saw a different situation where a similar condition was introduced, but where the requirement was that at most 5% of the population is included in the sample. Be careful not to confuse these two rules.

Example 9.5 Parents Tracking Kids Using GPS

Understand the context ➤

The report **"Parenting Children in the Age of Screens" (pewresearch.org, July 28, 2020)** summarized the results of a survey conducted by the Pew Research Center. In this survey, 1636 parents of children age 5 to 11 were asked if they track the location of their child using GPS apps or software. Of these parents, 540 responded yes. The sample was selected in a way that makes it reasonable to regard this sample as representative of the population of parents of children age 5 to 11. This information can be used to construct an estimate of the proportion of all parents of children age 5 to 11 who track the location of their child using GPS apps or software.

For this sample

$$\hat{p} = \frac{540}{1636} = 0.33$$

Formulate a plan ➤

Because the sample size is less than 10% of the population size and $n\hat{p} = 540$ and $n(1-\hat{p}) = 1096$ are both greater than or equal to 10, the conditions necessary for appropriate use of the formula for a large-sample confidence interval are met.

A 90% confidence interval for p is then

Do the work ➤

$$\begin{aligned}\hat{p} \pm (z \text{ critical value})\sqrt{\frac{\hat{p}(1-\hat{p})}{n}} &= 0.33 \pm 1.645\sqrt{\frac{(0.33)(1-0.33)}{1636}} \\ &= 0.33 \pm (1.645)(0.01) \\ &= 0.33 \pm 0.02 \\ &= (0.31, 0.35)\end{aligned}$$

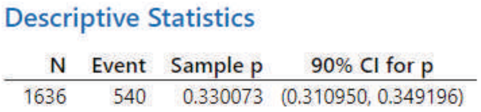
Interpret the results ➤

Based on these sample data, you can be 90% confident that the proportion of all parents of children age 5 to 11 who track the location of their child using GPS apps or software is between 0.31 and 0.35. You have used a *method* to construct this interval estimate that has a 10% error rate.

Statistical software or a graphing calculator can also be used to calculate a confidence interval for a population proportion. For example, the following output results from using Minitab

Copyright 2025 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s). Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

to calculate the 90% confidence interval for the data of Example 9.5. The small differences in the endpoints of the calculated interval are the result of rounding in the hand calculations.



Descriptive Statistics

N	Event	Sample p	90% CI for p
1636	540	0.330073	(0.310950, 0.349196)

The confidence level for the z confidence interval for a population proportion is only approximate. When you report a 95% confidence interval for a population proportion, the 95% confidence level implies that you used a method that produces an interval that includes the actual value of the population proportion 95% of the time in repeated random sampling. In fact, because the normal distribution is only an approximation to the sampling distribution of $\hat{p}$, the actual confidence level may differ somewhat from the reported value. If the conditions (1) $n\hat{p} \geq 10$ and $n(1 - \hat{p}) \geq 10$ and (2) n is at most 10% of the population size if sampling without replacement are met, the normal approximation is reasonable and the actual confidence level is usually not too different from the reported level. This is why it is important to check these conditions before calculating and reporting a large-sample z confidence interval for a population proportion.

What should you do if these conditions are not met? If the sample size is too small to satisfy the $n\hat{p}$ and $n(1 - \hat{p})$ greater than or equal to 10 condition, an alternative procedure can be used. One alternative method is the bootstrap confidence interval introduced in optional Section 9.5. If the condition that the sample size is less than 10% of the population size when sampling without replacement is not satisfied, the large-sample z confidence interval tends to be conservative (that is, it tends to be wider than is necessary to achieve the desired confidence level). In this case, a finite population correction factor can be used to obtain a more precise interval. Consult a statistician or a more advanced textbook for more information.

An Alternative to the Large-Sample *z* Interval

Investigators have shown that in some instances, even when the sample size conditions of the large-sample z confidence interval for a population proportion are met, the actual confidence level associated with the method may be noticeably different from the reported confidence level.

A modified interval that has an actual confidence level that is closer to the reported confidence level is based on a modified sample proportion, $\hat{p}_{\text{mod}}$, the proportion of successes after adding two successes and two failures to the sample. Then $\hat{p}_{\text{mod}}$ is

$$\hat{p}_{\text{mod}} = \frac{\text{number of successes} + 2}{n + 4}$$

$\hat{p}_{\text{mod}}$ is used in place of $\hat{p}$ in the usual confidence interval formula. Properties of this modified confidence interval are investigated in Activity 9.2 at the end of the chapter.

General Form of a Confidence Interval

Many confidence intervals have the same general form as the large-sample z confidence interval for p just considered. The interval was based on a point estimate $\hat{p}$. The standard deviation of this statistic is $\sqrt{p(1 - p)/n}$. This resulted in a confidence interval of the form

$$\begin{pmatrix}\text{point estimate using}\\ \text{a specified statistic}\end{pmatrix} \pm (\text{critical value})\begin{pmatrix}\text{standard deviation}\\ \text{of the statistic}\end{pmatrix}$$

Because p was unknown, the standard deviation of the statistic was estimated by $\sqrt{\hat{p}(1 - \hat{p})/n}$, which yielded an interval of the form

$$\begin{pmatrix}\text{point estimate using}\\ \text{a specified statistic}\end{pmatrix} \pm (\text{critical value})\begin{pmatrix}\text{estimated}\\ \text{standard deviation}\\ \text{of the statistic}\end{pmatrix}$$

For a population characteristic other than p, a statistic for estimating the characteristic is selected. Then (drawing on statistical theory) a formula for the standard deviation of the statistic is determined. In practice, it is almost always necessary to estimate this standard deviation, so the confidence interval has the following form:

$$\begin{pmatrix}\text{point estimate using}\\ \text{a specified statistic}\end{pmatrix} \pm (\text{critical value})\begin{pmatrix}\text{estimated}\\ \text{standard deviation}\\ \text{of the statistic}\end{pmatrix}$$

It is common practice to refer to both the standard deviation of a statistic and the *estimated* standard deviation of a statistic as the **standard error**. In this textbook, when the term standard error is used, it refers to the estimated standard deviation of a statistic.

Definition

Standard error: The estimated standard deviation of a statistic.

The 95% confidence interval for p is based on the fact that, for approximately 95% of all random samples, $\hat{p}$ is within $1.96\sqrt{\frac{p(1-p)}{n}}$ of p. The quantity $1.96\sqrt{\frac{p(1-p)}{n}}$ is sometimes called the **margin of error** associated with a 95% confidence level. This means that you are 95% confident that the point estimate $\hat{p}$ is no farther than this quantity from p.

Definition

Margin of error: If the sampling distribution of a statistic is (at least approximately) normal, the **margin of error, *M***, associated with a 95% confidence interval is (1.96)·(standard error of the statistic).

Choosing the Sample Size

Before collecting any data, an investigator may wish to determine a sample size for which a particular value of the margin of error is achieved. For example, with p representing the actual proportion of students at a university who purchase textbooks online, the objective of an investigation may be to estimate p to within 0.05 with 95% confidence.

The value of n necessary to achieve this is obtained by setting $1.96\sqrt{\frac{p(1-p)}{n}}$ equal to 0.05 and solving for n.

In general, suppose that you want to estimate p to within an amount M (the specified margin of error) with 95% confidence. To find the necessary sample size, consider the equation

$$M = 1.96\sqrt{\frac{p(1-p)}{n}}$$

Solving this equation for n results in

$$n = p(1-p)\left(\frac{1.96}{M}\right)^2$$

Unfortunately, the use of this formula requires the value of p, which is unknown. One possible way to proceed is to carry out a preliminary study and use the resulting data to get a rough estimate of p. In other cases, prior knowledge may suggest a reasonable estimate of p.

If there is no reasonable basis for estimating p and a preliminary study is not feasible, a conservative solution follows from the observation that $p(1-p)$ is never larger than 0.25 (its value when $p = 0.5$). Replacing $p(1-p)$ with 0.25, the maximum value, yields

$$n = 0.25\left(\frac{1.96}{M}\right)^2$$

Copyright 2025 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s). Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

Using this formula to obtain n gives you a sample size for which you can be 95% confident that $\hat{p}$ will be within M of p, no matter what the value of p.

> The sample size required to estimate a population proportion p to within an amount M with 95% confidence is
>
> $$n = p(1-p)\left(\frac{1.96}{M}\right)^2$$
>
> The value of p may be estimated using prior information. In the absence of any such information, using $p = 0.5$ in this formula gives a conservatively large value for the required sample size (this value of p gives a larger n than would any other value).

Example 9.6 Sniffing Out Cancer

Understand the context ➤ Dogs have a sense of smell that is much more powerful than humans. Because of this, dogs can be trained to identify the presence of odors unique to specific types of cancer. The article **"Meet the Dogs Who Can Sniff Out Cancer Better Than Some Lab Tests" (cnn.com, February 3, 2016)** describes a large-scale study that was being planned by researchers to discern if dogs really can identify the presence of cancer by sniffing urine samples.

Suppose you want to collect data that would allow you to estimate the long-run proportion of accurate identifications for a particular dog that has completed training. The dog has been trained to lie down when presented with a urine specimen from a patient with cancer and to remain standing when presented with a specimen from a person who does not have cancer. How many different urine specimens should be used if you want to estimate the long run proportion of correct identifications for this dog to within 0.05 with 95% confidence?

Do the work ➤ Using a conservative value of $p = 0.5$ in the formula for required sample size gives

$$n = p(1-p)\left(\frac{1.96}{M}\right)^2 = (0.5)(0.5)\left(\frac{1.96}{0.05}\right)^2 = 384.16$$

Interpret the results ➤ This means that a sample of at least 385 urine specimens should be used. Notice that in sample size calculations, you always round up.

Exercises 9.10–9.35

9.10 Explain which would result in a wider large-sample confidence interval for p: a 90% confidence level or a 95% confidence level. (Hint: Consider the confidence interval formula.)

9.11 Explain which would result in a wider large-sample confidence interval for p: a sample size of $n = 100$ or $n = 400$?

9.12 The formula used to calculate a large-sample confidence interval for p *is*

$$\hat{p} \pm (z \text{ critical value})\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

What is the appropriate z critical value for each of the following confidence levels?

a. 95%
b. 90%
c. 99%
d. 80%
e. 85%

9.13 The use of the interval

$$\hat{p} \pm (z \text{ critical value})\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

requires a large sample. For each of the following combinations of n and $\hat{p}$, indicate whether the sample size is large enough for use of this interval to be appropriate.

Copyright 2025 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s). Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

a. $n = 50$ and $\hat{p} = 0.30$
b. $n = 50$ and $\hat{p} = 0.05$
c. $n = 15$ and $\hat{p} = 0.45$
d. $n = 100$ and $\hat{p} = 0.01$

9.14 For each of the following combinations of sample size and sample proportion, indicate whether the sample size is large enough for appropriate use of the large-sample confidence interval for a population proportion.
a. $n = 100$ and $\hat{p} = 0.70$
b. $n = 40$ and $\hat{p} = 0.25$
c. $n = 60$ and $\hat{p} = 0.25$
d. $n = 80$ and $\hat{p} = 0.10$

9.15 Discuss how each of the following factors affects the width of the confidence interval for p. (Hint: Consider the confidence interval formula.)
a. The confidence level
b. The sample size
c. The value of $\hat{p}$

9.16 It might surprise you to learn that many employers use social media background checks as part of their evaluation of job candidates. In a survey of 500 hiring managers, 395 said that they have chosen not to hire a candidate due to inappropriate content on social media **("90% of Employers Consider an Applicant's Social Media Activity During Hiring Process," smallbiztrends.com, May 5, 2020)**. Suppose that this sample is representative of the population of hiring managers in the United States.
a. Use the given information to estimate the proportion of hiring managers who have chosen not to hire a job candidate due to inappropriate content on social media using a 95% confidence interval. Give an interpretation of the interval in context and an interpretation of the confidence level of 95%.
b. Would a 90% confidence interval be wider or narrower than the 95% confidence interval from Part (a)? Explain.

9.17 The title of the article **"40 Percent of Americans Think They Could Compete in the Olympics" (onepoll.us, August 7, 2021)** is based on data from a representative sample of 1000 adult Americans. The title implies that for this sample $\hat{p} = 0.40$, and this sample proportion could be used as an estimate of the population proportion of all adult Americans who think they could compete in the Olympics. Based on the sample data, a 90% confidence interval for the proportion of adult Americans who think they could compete in the Olympics is (0.375, 0.425). What is the meaning of the 90% confidence level associated with this interval?

9.18 The report **"Parents, Teens and Digital Monitoring" (Pew Research Center, January 7, 2016)** reported that 61% of parents of teens age 13 to 17 said that they had checked which websites their teens had visited. The 61% figure was based on a representative sample of 1060 parents of teens in this age group.
a. What assumption must be made in order for it to be appropriate to use the formula of this section to construct a confidence interval to estimate the proportion of all parents of teens age 13 to 17 who have checked which websites their teens visit?
b. Construct and interpret a 90% confidence interval for the proportion of all parents of teens age 13 to 17 who have checked which websites their teens visit. (Hint: See Example 9.5.)
c. Would a 99% confidence interval be narrower or wider than the interval calculated in Part (b)? Justify your answer.

9.19 The report **"The 2020 Consumer Financial Literacy Survey" (The National Foundation for Credit Counseling, March 23, 2020)** summarized data from a representative sample of 2004 adult Americans. Based on data from this sample, it was reported that over half of adult Americans would give themselves a grade of A or B on their knowledge of personal finance. This statement was based on observing that 1142 people in the sample would have given themselves a grade of A or B.
a. Construct and interpret a 95% confidence interval for the proportion of all adult Americans who would give themselves a grade of A or B on their knowledge of personal finance.
b. Is the confidence interval from Part (a) consistent with the statement that a majority of adult Americans would give themselves a grade of A or B? Explain why or why not.

9.20 The *USA TODAY Snapshot* titled **"Big Bang Theory" (*USA TODAY*, October 14, 2016)** summarized data from a sample of 1003 American parents of children age 6 to 11. It reported that 53% of these parents view science-oriented TV shows as a good way to expose their children to science outside of school. Assume that this sample is representative of American parents of children age 6 to 11. Construct and interpret a 90% confidence interval for the proportion of American parents of children age 6 to 11 who view science-oriented TV shows as a good way to expose their children to science.

9.21 The article **"Most Dog Owners Take More Pictures of Their Pet Than Their Spouse" (news.fastcompany.com, August 22, 2016)** indicates that in a sample of 1000 dog owners, 650 said that they take more pictures of their dog than their significant others or friends. In addition, 460 said that they are more likely to complain to their dog than to a friend. Suppose that it is reasonable to consider this sample as representative of the population of dog owners.

a. Construct and interpret a 90% confidence interval for the proportion of dog owners who take more pictures of their dog than of their significant others or friends.

b. Construct and interpret a 95% confidence interval for the proportion of dog owners who are more likely to complain to their dog than to a friend.

c. Give two reasons why the confidence interval in Part (b) is wider than the interval in Part (a).

9.22 The ***Princeton Review 2021 College Hopes and Worries Survey Report*** **(princetonreview.com)** reported that 32% of students applying to college wanted to attend a college that was within 250 miles of their home and that 49% of parents of students applying to college wanted their child to attend a college that was within 250 miles from home. The reported percentages were based on random samples of 11,113 students applying to college and of 2960 parents of students applying to college.

a. Construct and interpret a 90% confidence interval for the proportion of students applying to college who want to attend a college within 250 miles of their home.

b. Construct and interpret a 90% confidence interval for the proportion of parents of students applying to college who want their child to attend a college within 250 miles from home.

c. Explain why the two 90% confidence intervals are not the same width.

9.23 The report **"Job Seeker Nation Study 2016" (jobvite.com)** summarized a survey of 2305 working adults. The report indicates that 484 of the working adults surveyed said they were very concerned that their job will be automated, outsourced, or otherwise made obsolete in the next 5 years. The sample was selected in a way designed to produce a representative sample of working adults. Construct and interpret a 95% confidence interval for the proportion of working adults who are very concerned that their job will be automated, outsourced, or otherwise made obsolete in the next 5 years.

9.24 The article **"Write It by Hand to Make It Stick" (*Advertising Age*, July 27, 2016)** summarizes data from a survey of 1001 students of age 13 to 19 years. Of the students surveyed, 851 reported that they learn best using a mix of digital and nondigital tools. Construct and interpret a 95% confidence interval for the proportion of students of age 13 to 19 who would say that they learn best using a mix of digital and nondigital tools. In order for the method used to construct the interval to be valid, what assumption about the sample must be reasonable?

9.25 The article referenced in the previous exercise also indicated that 811 of the 1001 students surveyed said that they would feel restricted if they could only work on digital devices. Would a 95% confidence interval for the proportion of students of age 13 to 19 years who would say that they would feel restricted if they could only use digital devices be narrower or wider than the interval constructed in the previous exercise for the proportion who would say that they learn best using a mix of digital and nondigital tools? Explain your reasoning—you should be able to answer this question without constructing the second confidence interval.

9.26 The *USA TODAY Snapshot* titled **"Have a Nice Trip" (*USA TODAY*, November 17, 2015)** summarized data from a survey of 1000 adult Americans who had traveled by air at least once in the previous year. The *Snapshot* includes the following statement: "38% admit to yelling at a complete stranger while traveling." Suppose that the sample was selected to be representative of the population of adults who have traveled by air at least once in the previous year. Calculate and interpret a 95% confidence interval for the population proportion who have yelled at a complete stranger while traveling.

9.27 *Business Insider* reported that a study commissioned by eBay motors found that nearly 40% of millennials who drive a car that is more than 5 years old have named their cars **("Millennials Have an Odd Habit When It Comes to Their Cars," April 14, 2016)**. Suppose that this statement was based on a sample of size 800 and that 312 reported that they had named their car. Assuming that the sample was selected to be representative of the population of millennials who drive a car that is more than 5 years old, calculate and interpret a 90% confidence interval for the proportion of all millennials who drive a car that is more than 5 years old who have named their cars.

9.28 The report **"Global Trustworthiness Index 2021" (Ipsos.com, October 12, 2021)** summarizes data from a large international study of public opinion in 28 countries. The sample size was approximately 1000 for each country. Of the survey participants from the United States, 62% said that they thought scientists were trustworthy.

a. Assuming that the sample was selected to be representative of the population of adult Americans and that the sample size was 1000, what is an estimate of the population proportion who think that scientists are trustworthy?

b. Calculate and interpret the margin of error associated with your estimate in Part (a).

9.29 The article **"Most Americans Don't Understand the Cloud, But They Should" (foxbusiness.com, October 17, 2016)** reported that in a sample of 1000 people, 22% said they have pretended to know what the cloud is or how it works. If it is reasonable to regard the sample as representative of adult Americans, an estimate of the proportion who have pretended to know what the cloud is or how it works is 0.22. The margin of error associated with this estimate is 0.026. Interpret the value of this margin of error.

9.30 The Gallup Organization conducts an annual survey on crime. It was reported that 25% of all households experienced some sort of crime during the past year. This estimate was based on a sample of 1002 randomly selected households. The report states, "One can say with 95% confidence that the margin of sampling error is ±3 percentage points." Explain how this statement can be justified.

9.31 The article **"Paranormal Nation: Nearly Two-Thirds of Americans Believe in Ghosts or Aliens" (studyfinds.org, October 26, 2021)** summarizes data from a survey of 2000 adult Americans. Based on the survey data, the authors estimated that the proportion who believe in visitors from other planets is 0.39 and that the proportion who believe in Bigfoot is 0.27. Assuming that the sample can be viewed as a random sample of adult Americans, which of the two estimates (visitors from other planets or bigfoot) has a smaller margin of error? Provide a justification for your choice.

9.32 The article **"Half of Americans Prioritize Protecting the Environment Over the Economy" (YouGovAmerica, November 19, 2021)** summarized a survey of 1500 adult U.S. citizens. When asked if they expect to feel the effects of climate change in their lifetime, 885 of the survey participants responded yes. The article states that participants were selected to be representative of all adult U.S. citizens.

a. Estimate the proportion of adult U.S. citizens who think that they will feel the effects of climate change in their lifetime. What statistic did you use?

b. Use the sample data to estimate the standard error of $\hat{p}$.

c. Calculate and interpret the margin of error associated with the estimate in Part (a).

9.33 A discussion of digital ethics appears in the article **"Academic Dishonesty Among University Students" (*PLOS ONE*, August 21, 2020)**. One question of interest to people studying academic dishonesty is: What proportion of college students have used cell phones to cheat on an exam? Suppose you have been asked to estimate this proportion for students enrolled at a large college. How many students should you include in your sample if you want to estimate this proportion with a margin of error of 0.02?

9.34 The article **"Uncovering Password Habits: Are Users' Password Security Habits Improving?" (Data Insider, September 29, 2020)** estimated the proportion of Google users who remember their passwords by writing then on a piece of paper to be 0.386. The margin of error associated with this estimate is 0.034.

a. Suppose that you would like to estimate the proportion of Google users who remember their passwords by writing them on a piece of paper with a margin of error of 0.01. Would you want a sample size that was smaller or larger than the one that produced a margin of error of 0.034?

b. Using 0.386 as a preliminary estimate, what is the required sample size if you want to estimate the proportion of Google users who remember their passwords by writing them on a piece of paper with a margin of error of 0.01?

c. How does the sample size in Part (b) compare to the sample size that would result from using the conservative value of 0.5?

d. What sample size would you recommend? Justify your answer.

9.35 Many cars now have Wi-Fi capability, but the cost of a car Wi-Fi plan is estimated to be between \$20 and \$40 per month **(*Car and Driver*, July 10, 2020)**. Suppose that a small preliminary survey of new car buyers estimated the proportion who would be willing to pay between \$20 and \$40 per month to be around 0.30. Use the given preliminary estimate to determine the sample size required to estimate this proportion with a margin of error of 0.02.

Section 9.3 Confidence Interval for a Population Mean

In this section, you will see how to use information from a random sample to construct a confidence interval estimate of a population mean, μ. First, consider the case in which σ, the population standard deviation, is known and the sample size n is large enough for the Central Limit Theorem to apply. In this case, the following three properties about the sampling distribution of $\bar{x}$ hold:

1. The sampling distribution of $\bar{x}$ is centered at μ, so $\bar{x}$ is an unbiased statistic for estimating μ ($\mu_{\bar{x}} = \mu$).
2. The standard deviation of $\bar{x}$ is $\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$.
3. As long as n is large (generally $n \geq 30$), the sampling distribution of $\bar{x}$ is approximately normal, even when the population distribution itself is not normal.

The same reasoning that was used to develop the large-sample confidence interval for a population proportion p can be used to obtain a confidence interval estimate for μ.

The One-Sample z Confidence Interval for μ

The general formula for a confidence interval for a population mean μ when

1. $\bar{x}$ is the sample mean from a **simple random sample**,
2. the **sample size n is large** (generally $n \geq 30$), and
3. σ, **the population standard deviation, is known**

is given below.

$$\bar{x} \pm (z \text{ critical value})\left(\frac{\sigma}{\sqrt{n}}\right)$$

Example 9.7 Cosmic Radiation

Understand the context ➤

Cosmic radiation levels rise with increasing altitude, prompting researchers to consider how pilots and flight crews are affected by increased exposure to cosmic radiation. The National Council on Radiation Protection and Measurements estimates that flight crew members have a mean annual radiation exposure of 3.07 millisievert (mSv) per year **(Centers for Disease Control and Prevention, 2017).** Suppose that the estimated mean exposure for flight crew members is based on a random sample of 100 flight crew members.

Here, μ will represent the mean annual cosmic radiation exposure for all flight crew members. Although σ, the actual value of the population standard deviation, is not usually known, for purposes of this example, suppose that σ is known to be 0.35 mSv. Because the sample size is large and σ is known, a 95% confidence interval for μ is

z critical value for 95% confidence level

Do the work ➤

$$\bar{x} \pm (z \text{ critical value})\frac{\sigma}{\sqrt{n}} = 3.07 \pm \overset{\downarrow}{(1.96)}\frac{0.35}{\sqrt{100}}$$

$$= 3.07 \pm 0.069$$

$$= (3.001, 3.139)$$

Interpret the results ➤

Based on this sample, *plausible* values of μ, the mean annual cosmic radiation exposure for all flight crew members, are those between 3.001 and 3.139 mSv. One mSv is the average annual exposure in the United States due to normal exposure to background radiation. This means that it is estimated that the mean annual exposure of flight crew members is somewhere between about 3 and 3.14 times greater than for the general population. A confidence level of 95% is associated with the method used to produce this interval estimate.

The confidence interval just introduced is appropriate when σ is known and n is large, and it can be used regardless of the shape of the population distribution. This is because this confidence interval is based on the Central Limit Theorem, which says that when n is sufficiently large, the sampling distribution of $\bar{x}$ is approximately normal for any population distribution.

When n is small, the Central Limit Theorem cannot be used to justify the normality of the $\bar{x}$ sampling distribution, so the z confidence interval cannot be used. One way to proceed in the small-sample case is to make a specific assumption about the shape of the population distribution and then to use a method that is valid under this assumption.

One instance where this is easy to do is when the population distribution is normal in shape. Recall that for a normal population distribution, the sampling distribution of $\bar{x}$ is normal even for small sample sizes. So, if n is small but the population distribution is normal, the same confidence interval formula just introduced can still be used.

If it is reasonable to believe that the distribution of values in the population is normal, a confidence interval for μ (when σ is known) is

$$\bar{x} \pm (z \text{ critical value})\left(\frac{\sigma}{\sqrt{n}}\right)$$

This interval is appropriate even when n is small, as long as it is reasonable to think that the population distribution is normal in shape.

There are several ways that sample data can be used to assess the plausibility of normality. Two common ways are to look at a normal probability plot of the sample data (looking for a plot that is reasonably straight) or to construct a dotplot or a boxplot of the data (looking for approximate symmetry and no outliers).

Confidence Interval for μ When σ Is Unknown

The confidence interval just developed has an obvious drawback: To calculate the interval endpoints, σ must be known. Unfortunately, this is rarely the case in practice. Next, consider the situation when σ is unknown. The development of the confidence interval in this instance depends on the assumption that the population distribution is normal. *This assumption is not critical if the sample size is large, but it is important when the sample size is small.*

To understand the derivation of this confidence interval, begin by taking another look at the 95% confidence interval when σ is known. You know that $\mu_{\bar{x}} = \mu$ and $\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$. Also, when the population distribution is normal, the $\bar{x}$ distribution is normal. These facts imply that the standardized variable

$$z = \frac{\bar{x} - \mu}{\frac{\sigma}{\sqrt{n}}}$$

has approximately a standard normal distribution. Because the interval from -1.96 to 1.96 captures an area of 0.95 under the z curve, approximately 95% of all samples result in an $\bar{x}$ value that satisfies

$$-1.96 < \frac{\bar{x} - \mu}{\frac{\sigma}{\sqrt{n}}} < 1.96$$

Copyright 2025 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s). Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

Rewriting these inequalities to isolate μ in the middle results in the equivalent inequalities:

$$\bar{x} - 1.96\left(\frac{\sigma}{\sqrt{n}}\right) < \mu < \bar{x} + 1.96\left(\frac{\sigma}{\sqrt{n}}\right)$$

The term $\bar{x} - 1.96\left(\frac{\sigma}{\sqrt{n}}\right)$ is the lower endpoint of the 95% large-sample confidence interval for μ, and $\bar{x} + 1.96\left(\frac{\sigma}{\sqrt{n}}\right)$ is the upper endpoint.

If σ is unknown, you can use the sample data to estimate σ. If you use the sample standard deviation as your estimate, the result is a different standardized variable denoted by t:

$$t = \frac{\bar{x} - \mu}{\frac{s}{\sqrt{n}}}$$

The value of s may not be all that close to σ, especially when n is small. As a consequence, the use of s in place of σ introduces extra variability. The value of z varies from sample to sample, because different samples generally result in different $\bar{x}$ values. There is even more variability in t, because different samples may result in different values of both $\bar{x}$ and s. Because of this, the distribution of t is more spread out than the standard normal (z) distribution.

Developing an appropriate confidence interval requires investigating the probability distribution of the standardized variable t for a sample from a normal population. To do this, you must first learn about probability distributions called *t distributions*.

t Distributions

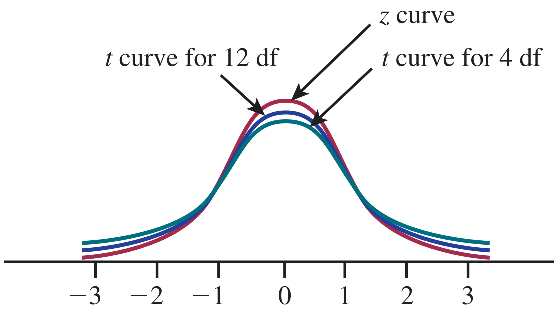
Just as there are many different normal distributions, there are also many different t distributions. While normal distributions are distinguished from one another by their mean μ and standard deviation σ, t distributions are distinguished by a positive whole number called the number of *degrees of freedom* (df). There is a t distribution with 1 df, another with 2 df, and so on.

Important Properties of *t* Distributions

1. The t distribution corresponding to any particular number of degrees of freedom is bell-shaped and centered at zero (just like the standard normal (z) distribution).
2. Each t distribution is more spread out than the standard normal (z) distribution.
3. As the number of degrees of freedom increases, the variability of the corresponding t distribution decreases.
4. As the number of degrees of freedom increases, the corresponding sequence of t distributions approaches the standard normal (z) distribution.

The properties discussed in the preceding box are illustrated in Figure 9.6, which shows two t curves along with the z curve.

Figure 9.6
Comparison of the z curve and t curves for 12 df and 4 df.

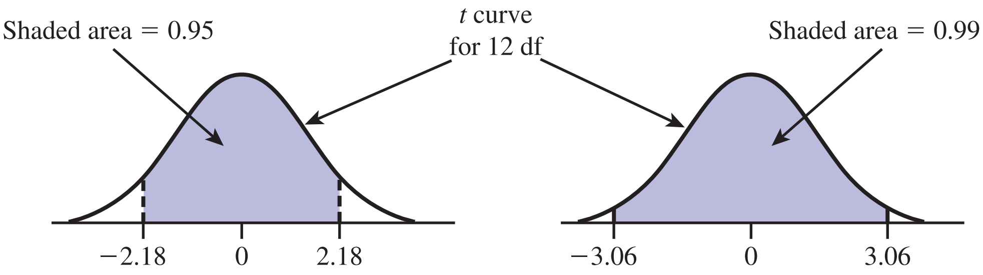


Copyright 2025 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s). Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

Appendix Table 3 gives selected critical values for various *t* distributions. The central areas for which values are tabulated are 0.80, 0.90, 0.95, 0.98, 0.99, 0.998, and 0.999. To find a particular critical value, go down the left margin of the table to the row labeled with the desired number of degrees of freedom. Then move over in that row to the column headed by the desired central area.

For example, the value in the 12-df row under the column corresponding to central area 0.95 is 2.18, so 95% of the area under the *t* curve with 12 df lies between −2.18 and 2.18. Moving over two columns, the critical value for central area 0.99 (still with 12 df) is 3.06 (see Figure 9.7). Moving down the 0.99 column to the 20-df row, the critical value is 2.85, so the area between −2.85 and 2.85 under the *t* curve with 20 df is 0.99.

Figure 9.7
t critical values illustrated.



Notice that the critical values increase from left to right in each row of Appendix Table 3. This makes sense because as you move to the right, you capture larger central areas. In each column, the critical values decrease as you move downward, reflecting decreasing variability for *t* distributions with larger degrees of freedom.

The larger the number of degrees of freedom, the more closely the *t* curve resembles the *z* curve. To emphasize this, the *z* critical values are included as the last row of the *t* table. Also notice that, once the number of degrees of freedom is greater than 30, the critical values change very little as the number of degrees of freedom increases. For this reason, Appendix Table 3 jumps from 30 df to 40 df, then to 60 df, then to 120 df, and finally to the row of *z* critical values.

If you need a critical value for a number of degrees of freedom between those tabulated, you can just use the critical value for the closest df. For df > 120, you can use the *z* critical values. Many graphing calculators and statistical software packages calculate *t* critical values for any number of degrees of freedom, so if you are using such a calculator, it is not necessary to approximate the *t* critical values as described.

One-Sample *t* Confidence Interval

The fact that the sampling distribution of $\dfrac{\bar{x} - \mu}{(\sigma/\sqrt{n})}$ is approximately the z (standard normal) distribution when n is large led to the z confidence interval when σ is known. In the same way, the following proposition allows you to obtain a confidence interval when the population distribution is normal but σ is unknown.

If $\bar{x}$ and s are the mean and standard deviation of a random sample from a normal population distribution, then the probability distribution of the standardized variable

$$t = \frac{\bar{x} - \mu}{\dfrac{s}{\sqrt{n}}}$$

is the t distribution with df $= n - 1$.

Copyright 2025 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s). Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

To see how this result leads to the desired confidence interval, consider the case $n = 25$. You would then use the t distribution with $\text{df} = n - 1 = 24$. From Appendix Table 3, the interval between -2.06 and 2.06 captures a central area of 0.95 under the t curve with 24 df. This means that 95% of all samples of size $n = 25$ from a normal population result in values of $\bar{x}$ and s for which

$$-2.06 < \frac{\bar{x} - \mu}{\frac{s}{\sqrt{n}}} < 2.06$$

Rewriting these inequalities to isolate μ yields

$$\bar{x} - 2.06\left(\frac{s}{\sqrt{25}}\right) < \mu < \bar{x} + 2.06\left(\frac{s}{\sqrt{25}}\right)$$

The 95% confidence interval for μ in this situation extends from the lower endpoint $\bar{x} - 2.06\left(\frac{s}{\sqrt{25}}\right)$ to the upper endpoint $\bar{x} + 2.06\left(\frac{s}{\sqrt{25}}\right)$. This interval can also be written

$$\bar{x} \pm 2.06\left(\frac{s}{\sqrt{25}}\right)$$

The differences between this interval and the interval when σ is known are the use of the t critical value 2.06 rather than the z critical value 1.96 and the use of the sample standard deviation as an estimate of σ. The extra uncertainty that results from estimating σ is why the t interval is wider than the z interval.

If the sample size is something other than 25 or if the desired confidence level is something other than 95%, a different t critical value (obtained from Appendix Table 3 or by using technology) is used in place of 2.06.

The One-Sample *t* Confidence Interval for μ

The general formula for a confidence interval for a population mean μ based on a sample of size n when

1. $\bar{x}$ is the sample mean from a **simple random sample**,
2. the **population distribution is normal**, *or* **the sample size n is large** (generally $n \geq 30$), and
3. σ, **the population standard deviation, is unknown**

is

$$\bar{x} \pm (t \text{ critical value})\left(\frac{s}{\sqrt{n}}\right)$$

where the t critical value is based on $\text{df} = n - 1$. Appendix Table 3 gives critical values appropriate for each of the confidence levels 90%, 95%, and 99%, as well as several other less frequently used confidence levels.

If n is large (generally $n \geq 30$), the normality of the population distribution is not critical. However, this confidence interval is appropriate for small n *only* when the population distribution is (at least approximately) normal. If this is not the case, as might be suggested by a normal probability plot or boxplot, a different method should be used (for example, the method introduced in Section 9.6).

Example 9.8 Drive-Through Medicine

Understand the context ➤

During a flu outbreak, many people visit emergency rooms, where they often must wait in crowded waiting rooms where other patients may be exposed. The paper **"Drive-Through Medicine: A Novel Proposal for Rapid Evaluation of Patients during an Influenza**

Pandemic" (*Annals of Emergency Medicine* [2010]: 268–273) describes an interesting study of the feasibility of a drive-through model where flu patients are evaluated while they remain in their cars. One of the interesting observations from this study was that not only were patients kept relatively isolated and away from other patients, but the time to process a patient was shorter because delays related to turning over examination rooms were eliminated.

In the study, 38 volunteers were each given a scenario from a randomly selected set of flu cases seen in the emergency room. The scenarios provided the volunteer with a medical history and a description of symptoms that would allow the volunteer to respond to questions from the examining physician. These volunteer patients were then processed using a drive-through procedure that was implemented in the parking structure of Stanford University Hospital. The time to process each case from admission to discharge was recorded.

Consider the data ➤

Data read from a graph that appears in the paper were used to calculate the following summary statistics for admission-to-discharge processing times (in minutes):

$$n = 38 \qquad \bar{x} = 26 \qquad s = 1.57$$

What Do You Think?

Are the conditions needed for the one-sample *t* confidence interval met for this example?

Formulate a plan ➤

A boxplot of the 38 processing times did show a couple of outliers on the high end, corresponding to unusually long processing times, suggesting that it is probably not reasonable to think of the population distribution of drive-through processing times as being approximately normal. However, because the sample size is greater than 30 and the distribution of sample processing times was not extremely skewed, it is appropriate to consider using the *t* confidence interval to estimate the mean admission-to-discharge processing time for flu patients using the drive-through procedure. Because the 38 flu scenarios were thought to be representative of the population of flu patients seen in emergency rooms and the sample size is large, you can use the formula for the *t* confidence interval to calculate a 95% confidence interval.

For this example, $n = 38$, df $= 37$, and the appropriate *t* critical value is 2.02 (from the 40-df row of Appendix Table 3). The 95% confidence interval is then

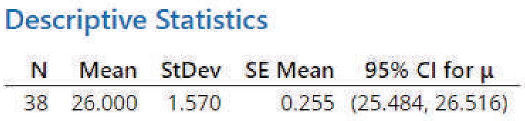
Do the work ➤

$$\begin{aligned}\bar{x} \pm (t \text{ critical value})\left(\frac{s}{\sqrt{n}}\right) &= 26 \pm (2.02)\left(\frac{1.57}{\sqrt{38}}\right)\\ &= 26 \pm 0.514\\ &= (25.486, 26.514)\end{aligned}$$

Interpret the results ➤

Based on the sample data, you would estimate that the actual mean admission-to-discharge processing time for flu patients processed using the drive-through procedure is between 25.486 minutes and 26.514 minutes. A method that has a 5% error rate was used to construct this interval. The authors of the paper indicated that the average processing time for flu patients seen in the emergency room was about 90 minutes, so it appears that the drive-through procedure has promise both in terms of keeping flu patients isolated and also in reducing processing time, on average.

Statistical software or a graphing calculator can be used to produce a one-sample *t* confidence interval. If you have access to technology, you probably will want to use it to do the calculations. For example, Minitab output for a 95% confidence interval for the mean processing time for emergency room flu cases using the new drive-through model is shown here.



```
Descriptive Statistics

 N    Mean   StDev  SE Mean   95% CI for μ
38  26.000   1.570    0.255  (25.484, 26.516)
```

Copyright 2025 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s). Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

The small difference between the confidence interval produced by Minitab and the one in Example 9.8 is just due to rounding in the hand calculations.

Example 9.9 Electric Car Energy Consumption

Understand the context ➤

● The Environmental Protection Agency introduced "miles per gallon-equivalent" (MPG-e) as a way to compare the amount of energy consumed by electric and other alternative fuel vehicles to gas-powered vehicles. MPG-e is a measure of how far an electric vehicle can travel on 33.7 kWh of electricity, which is the approximately equivalent to 1 gallon of gas. This means that a rating of 40 MPG-e for an electric car is comparable to a gas-powered car that gets 40 miles per gallon of gas.

Consider the data ➤

The **U.S. Department of Energy** reports MPG-e for electric vehicles sold in the United States **(fueleconomy.gov).** The accompanying data are the MPG-e ratings for a random sample of seven 2021 model electric cars.

75	92	96	113	129	111	104

You can use these data to estimate the mean MPG-e for the population that consists of all 2021 model electric cars.

Formulate a plan ➤

There are two conditions that must be met in order for the one-sample t confidence interval to be appropriate. It was stated that the sample was a random sample from the population of 2021 model electric cars. Because the sample size is small ($n = 7$), you need to consider whether it is reasonable to think that the distribution of MPG-e ratings for the population of all 2021 car models is at least approximately normal. Figure 9.8 is a dotplot and Figure 9.9 is a normal probability plot of the sample data. Although it is difficult to assess with only seven data values, the dotplot is approximately symmetric and the normal probability plot is reasonably straight, so it seems reasonable to think that the population distribution is at least approximately normal.

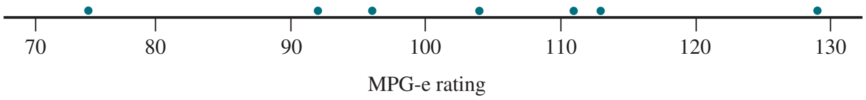


Figure 9.8
Dotplot for the sample data of Example 9.9

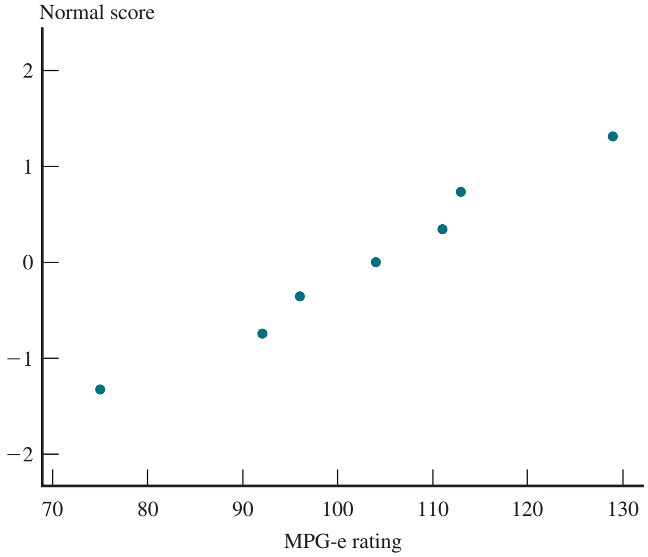


Figure 9.9
Normal probability plot for the sample data of Example 9.9

Do the work ➤

Calculation of a confidence interval for the mean MPG-e rating for the population of 2021 model electric cars requires $\bar{x}$ and s. From the given data,

$$\bar{x} = 102.86 \qquad s = 17.30$$

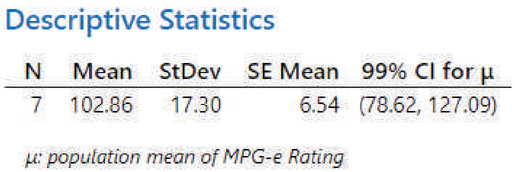
● **Data set available online**

Copyright 2025 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s). Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

The t critical value for a 99% confidence interval based on 6 df is 3.71 (obtained from Appendix Table 3 or using technology). The 99% confidence interval is

$$\begin{aligned}\bar{x} \pm (t \text{ critical value})\left(\frac{s}{\sqrt{n}}\right) &= 102.86 \pm (3.71)\left(\frac{17.30}{\sqrt{7}}\right) \\ &= 102.86 \pm 24.26 \\ &= (78.60, 127.12)\end{aligned}$$

A statistical software package or a graphing calculator could also have been used to calculate the end points of the confidence interval. The following output is from Minitab. The small difference between the confidence interval produced by Minitab and the one calculated by hand is just due to rounding in the hand calculations.



Descriptive Statistics

N	Mean	StDev	SE Mean	99% CI for μ
7	102.86	17.30	6.54	(78.62, 127.09)

μ: population mean of MPG-e Rating

Interpret the results ➤

Based on this sample, you can be 99% confident that the population mean MPG-e rating for 2021 car models is between 78.60 and 127.12. The 99% confidence level associated with this interval tells you that the method used to produce this estimate is successful in capturing the actual population mean about 99% of the time.

Choosing the Sample Size

When estimating μ using a large sample or using a small sample from a normal population, the margin of error, M, associated with a 95% confidence interval is

$$M = 1.96\left(\frac{\sigma}{\sqrt{n}}\right)$$

Before collecting any data, an investigator may wish to determine a sample size for which a particular value of the margin of error is achieved. For example, with μ representing the average fuel efficiency (in miles per gallon, mpg) for all cars of a certain type, the objective of an investigation may be to estimate μ to within 1 mpg with 95% confidence. The value of n necessary to achieve this is obtained by setting $M = 1$ and then solving

$$1 = 1.96\left(\frac{\sigma}{\sqrt{n}}\right)$$

for n.

In general, suppose that you want to estimate μ to within an amount M (the specified margin of error) with 95% confidence. Finding the necessary sample size requires solving the equation

$$M = 1.96\left(\frac{\sigma}{\sqrt{n}}\right)$$

for n. The result is

$$n = \left(\frac{1.96\sigma}{M}\right)^2$$

Notice that the greater the variability in the population (larger σ), the greater the required sample size will be. And, of course, the smaller the desired margin of error is, the larger the required sample size will be.

Copyright 2025 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s). Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

Use of the sample-size formula requires that σ be known, but this is rarely the case in practice. One possible strategy for estimating σ is to carry out a preliminary study and use the resulting sample standard deviation (or a somewhat larger value, to be conservative) to determine n for the main part of the study. Another possibility is simply to make an educated guess about the value of σ and to use that value to calculate n. For a population distribution that is not too skewed, the anticipated range (the difference between the largest and the smallest values) divided by 4 is sometimes used as a rough estimate of the value of the standard deviation.

The sample size required to estimate a population mean μ to within an amount M with 95% confidence is

$$n = \left(\frac{1.96\sigma}{M}\right)^2$$

If σ is unknown, it may be estimated based on previous information or, for a population that is not too skewed, by using (range)/4.

If the desired confidence level is something other than 95%, 1.96 is replaced by the appropriate z critical value (for example, 2.58 for 99% confidence).

Example 9.10 Cost of Textbooks

A college financial aid advisor wants to estimate the mean cost of textbooks per quarter for students at a particular university. For the estimate to be useful, it should be within $M = \$20$ of the actual population mean. How large a sample should be used to be 95% confident of achieving this level of accuracy?

To determine the required sample size, you need a value for σ. Suppose the advisor thinks that the amount spent on books varies widely, with most values between \$150 and \$550. A reasonable estimate of σ is then

$$\frac{range}{4} = \frac{550 - 150}{4} = \frac{400}{4} = 100$$

The required sample size is

$$n = \left(\frac{1.96\sigma}{M}\right)^2 = \left(\frac{(1.96)(100)}{20}\right)^2 = (9.8)^2 = 96.04$$

Rounding up, a sample size of 97 or larger is recommended.

Exercises 9.36–9.54

● Data set available online

9.36 Given a variable that has a t distribution with the specified degrees of freedom, what percentage of the time will its value fall in the indicated region?

a. 10 df, between -1.81 and 1.81
b. 10 df, between -2.23 and 2.23
c. 24 df, between -2.06 and 2.06
d. 24 df, between -2.80 and 2.80
e. 24 df, outside the interval from -2.80 to 2.80
f. 24 df, to the right of 2.80
g. 10 df, to the left of -1.81

9.37 The formula used to calculate a confidence interval for the mean of a normal population when n is small is

$$\bar{x} \pm (t \text{ critical value})\frac{s}{\sqrt{n}}$$

What is the appropriate t critical value for each of the following confidence levels and sample sizes?

a. 95% confidence, $n = 17$
b. 90% confidence, $n = 12$

Copyright 2025 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s). Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

c. 99% confidence, $n = 24$
d. 90% confidence, $n = 25$
e. 90% confidence, $n = 13$
f. 95% confidence, $n = 10$

9.38 The two intervals (114.4, 115.6) and (114.1, 115.9) are confidence intervals (calculated using the same sample data) for μ = mean weight of red delicious apples (in grams).

a. What is the value of the sample mean weight? (Hint: Where is the confidence interval centered?)
b. The confidence level for one of these intervals is 90% and for the other it is 99%. Which is which, and how can you tell?

9.39 Samples of two different models of cars were selected, and the actual speed for each car was determined when the speedometer registered 50 mph. The resulting 95% confidence intervals for mean actual speed were (51.3, 52.7) and (49.4, 50.6). Assuming that the two sample standard deviations are equal, which confidence interval is based on the larger sample size? Explain your reasoning.

9.40 The paper **"Relationship Between Food Waste, Diet Quality, and Environmental Sustainability" (*PLOS ONE*, April 18, 2018)** summarizes data from a study of food waste in the United States. Data from a representative sample of approximately 500 adult Americans resulted in a mean food waste per day of 422 grams. For purposes of this exercise, assume that the sample size was 500 and that the sample standard deviation was 530 grams (this is consistent with summary values given in the paper). Construct and interpret a 98% confidence interval for the mean food waste weight for adult Americans.

9.41 The paper **"The Effects of Adolescent Volunteer Activities on the Perception of Local Society and Community Spirit Mediated by Self-Conception" (*Advanced Science and Technology Letters* [2016]: 19–23)** describes a survey of a large representative sample of middle school children in South Korea. One question in the survey asked how much time per year the children spent in volunteer activities. The sample mean was 14.76 hours and the sample standard deviation was 16.54 hours.

a. Based on the reported sample mean and sample standard deviation, explain why it is not reasonable to think that the distribution of volunteer times for the population of South Korean middle school students is approximately normal.
b. The sample size was not given in the paper, but the sample size was described as large. Suppose that the sample size was 500. Explain why it is reasonable to use a one-sample t confidence interval to estimate the population mean even though the population distribution is not approximately normal.
c. Calculate and interpret a confidence interval for the mean number of hours spent in volunteer activities per year for South Korean middle school children.

9.42 ● Medical research has shown that repeated wrist extensions beyond 20 degrees increase the risk of wrist and hand injuries. Each of 24 students at Cornell University used a proposed new computer mouse design, and while using the mouse, each student's wrist extension was recorded. Data consistent with summary values given in the paper **"Comparative Study of Two Computer Mouse Designs" (Cornell Human Factors Laboratory Technical Report RP7992)** are given here.

27	28	24	26	27	25	25	24	24	24	25	28
22	25	24	28	27	26	31	25	28	27	27	25

a. Use these data to estimate the mean wrist extension for people using this new mouse design using a 90% confidence interval.
b. What assumptions are required in order for it to be appropriate to generalize your estimate to the population of Cornell students? To the population of all university students?
c. Based on your interval from Part (a), do you think there is reason to believe that the mean wrist extension for people using the new mouse design is greater than 20 degrees? Explain why or why not.

9.43 Students in a representative sample of 65 first-year students selected from a large university participated in a study of academic procrastination **("Study Goals and Procrastination Tendencies at Different Stages of the Undergraduate Degree," *Studies in Higher Education* [2016]: 2028–2043)**. Each student in the sample completed the Tuckman Procrastination Scale, which measures procrastination tendencies. Scores on this scale can range from 16 to 64, with scores over 40 indicating higher levels of procrastination. For the 65 first-year students in this study, the mean score on the procrastination scale was 37.0 and the standard deviation was 6.44.

a. Construct a 95% confidence interval estimate of μ, the mean procrastination scale for first-year students at this college.
b. Based on your interval, is 40 a plausible value for the population mean score? What does this imply about the population of first-year students?

9.44 The paper referenced in the previous exercise also reported that for a representative sample of 68 second-year students at the university, the sample mean procrastination score was 41.0 and the sample standard deviation was 6.82.

a. Construct a 95% confidence interval estimate of μ, the population mean procrastination score for second-year students at this college.

b. How does the confidence interval for the population mean score for second-year students compare to the confidence interval for first-year students calculated in the previous exercise? What does this tell you about the difference between first-year and second-year students in terms of mean procrastination score?

9.45 Suppose that a random sample of 50 bottles of a particular brand of liquid cough medicine is selected and the alcohol content of the liquid in each bottle is determined. Let μ denote the mean alcohol content (in percent) for the population of all bottles of the brand under study. Suppose that the sample of 50 results in a 95% confidence interval for μ of (7.8, 9.4).

a. Would a 90% confidence interval have been narrower or wider than the given interval? Explain your answer.

b. Consider the following statement: There is a 95% chance that μ is between 7.8 and 9.4. Is this statement correct? Why or why not?

c. Consider the following statement: If the process of selecting a random sample of size 50 and then computing the corresponding 95% confidence interval is repeated 100 times, 95 of the resulting intervals will include μ. Is this statement correct? Why or why not?

9.46 The authors of the paper **"Driving Performance While Using a Mobile Phone: A Simulation Study of Greek Professional Drivers" (*Transportation Research Part F* [2016]: 164–170)** describe a study to evaluate the effect of mobile phone use by taxi drivers in Greece. Fifty taxi drivers drove in a driving simulator where they were following a lead car. The drivers were asked to carry out a conversation while driving, and the following distance (the distance between the taxi and the lead car) was recorded. The sample mean following distance was 3.2 meters and the sample standard deviation was 1.11 meters.

a. Construct and interpret a 95% confidence interval for μ, the population mean following distance while talking on a mobile phone for the population of taxi drivers.

b. What assumption must be made to generalize this confidence interval to the population of all taxi drivers in Greece?

9.47 The article **"A Quasi-Experimental Study of the Impact of School Start Time Changes on Adolescent Sleep" (*Sleep Health* [2017]: 437–443)** describes a study in which a representative sample of high school students with a school start time of 7:20 a.m. and a representative sample of high school students with a school start time of 8:10 a.m. were surveyed about how much sleep (in minutes) they get on nights when there is school the next day. Summary statistics given in the article were used to calculate the following confidence intervals.

School Start Time	95% Confidence Interval
7:20 a.m.	(441.4, 448.7)
8:10 a.m.	(463.8, 474.2)

a. Suppose that the sample sizes for both samples were equal. Based on the given confidence intervals, which of the start time samples had the greatest variability in sleep time? Explain your choice.

b. Now suppose that the sample standard deviations for both samples were equal, but that the sample sizes might have been different. Which of the start time samples had the larger sample size? Explain your choice.

c. The interval (462.2, 475.8) is either a 90% confidence interval or a 99% confidence interval for the mean sleep time calculated using the sample data for 8:10 a.m. school start time. Is the confidence level for this interval 90% or 99%? Explain your choice.

9.48 The authors of the paper **"Length of Stay, Wait Time to Surgery and 30-Day Mortality for Patients with Hip Fractures After Opening of a Dedicated Orthopedic Weekend Trauma Room" (*Canadian Journal of Surgery* [2016]: 337–341)** were interested in estimating the mean time that patients who broke a hip had to wait for surgery before and after the opening of a new hospital facility. They reported that for a representative sample of 405 patients seen at this hospital before the new facility was opened, the mean wait time for surgery was 31.5 hours and the standard deviation of wait times was 27.0 hours. For a representative sample of 204 patients seen after the new facility was opened, the mean wait time for surgery was 28.5 hours and the standard deviation of wait times was 16.8 hours.

a. Construct and interpret a 95% confidence interval for the mean wait time of patients seen before the new facility was opened.

b. Construct and interpret a 95% confidence interval for the mean wait time of patients seen after the new facility was opened.

9.49 The article **"Americans' Big Debt Burden Growing, Not Evenly Distributed" (www.gallup.com, February 4, 2016)** reported that for a representative sample of Americans born between 1965 and 1971 (known as Generation X), the sample mean number of credit cards owned was 4.5. Suppose that the sample standard deviation (which was not reported) was 1.0 and that the sample size was $n = 300$.

a. Construct a 95% confidence interval for the population mean number of credit cards owned by Generation X Americans.

b. The interval in Part (a) does not include 0. Does this imply that all Generation X Americans have at least one credit card? Explain.

9.50 ● Example 9.3 gave the following airborne times (in minutes) for 10 randomly selected flights from San Francisco to Washington Dulles airport:

270 256 267 285 274 275 266 258 271 281

a. Calculate and interpret a 90% confidence interval for the mean airborne time for flights from San Francisco to Washington Dulles. (Hint: See Example 9.9.)

b. Give an interpretation of the 90% confidence level associated with the interval estimate in Part (a).

c. If a flight from San Francisco to Washington Dulles is scheduled to depart at 10 a.m., what would you recommend for the published arrival time? Explain.

9.51 The report **"NACE Salary Survey Executive Summary" (National Association of Colleges and Employers, Winter 2022)** states that the mean yearly salary for students graduating with mathematics and science degrees in 2021 was \$63,316. Suppose that a random sample of 50 mathematics and science graduates at a large university who received job offers resulted in a mean offer of \$64,000 and a standard deviation of \$3,300.

a. Use a 95% confidence interval to estimate the mean salary offer for students graduating with mathematics or science degrees from this university.

b. Based on your interval from Part (a), do you think that it is reasonable to conclude that the mean salary offer for students graduating with mathematics or science degrees from this university is different from the reported national average of \$63,316? Explain the reasoning you used to make this decision.

9.52 ● Five students visiting the student health center for a free dental examination during National Dental Hygiene Month were asked how many months had passed since their last visit to a dentist. Their responses were as follows:

6 17 11 22 29

Assuming that these five students can be considered a random sample of all students participating in the free checkup program, construct a 95% confidence interval for the mean number of months elapsed since the last visit to a dentist for the population of students participating in the program.

9.53 The Bureau of Alcohol, Tobacco, and Firearms (BATF) has been concerned about lead levels in California wines. In a previous testing of wine specimens, lead levels ranging from 50 to 700 parts per billion were recorded. How many wine specimens should be tested if the BATF wishes to estimate the mean lead level for California wines to within 10 parts per billion with 95% confidence? (Hint: See Example 9.10.)

9.54 The formula described in this section for determining sample size corresponds to a confidence level of 95%. What would be the appropriate formula for determining sample size when the desired confidence level is 90%? 98%?

Section 9.4 Interpreting and Communicating the Results of Statistical Analyses

The purpose of most surveys and many research studies is to produce estimates of population characteristics. One way of providing an estimate is to construct and report a confidence interval for the population characteristic of interest.

Communicating the Results of Statistical Analyses

When using sample data to estimate a population characteristic, a point estimate or a confidence interval estimate might be used. Confidence intervals are generally preferred because a point estimate by itself does not convey any information about the accuracy of the estimate. For this reason, whenever you report the value of a point estimate, it is a good idea to also include an estimate of the margin of error.

Reporting and interpreting a confidence interval estimate requires a bit of care. First, always report both the confidence interval and the confidence level associated with the

Copyright 2025 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s). Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

method used to produce the interval. Then, remember that both the confidence interval and the confidence level should be interpreted.

A good strategy is to begin with an interpretation of the confidence interval in the context of the problem and then to follow that with an interpretation of the confidence level. For example, if a 90% confidence interval for p, the proportion of students at a particular university who own a laptop computer, is (0.56, 0.78), you might say

interpretation of interval → I am 90% confident that between 56% and 78% of the students at this university own laptops.

explanation of "90% confidence" interpretation of confidence level → I used a method to produce this estimate that is successful in capturing the actual population proportion 90% of the time.

When providing an interpretation of a confidence interval, remember that the interval is an estimate of a population characteristic and be careful *not* to say that the interval applies to individual values in the population or to the values of sample statistics. For example, if a 99% confidence interval for μ, the mean amount of ketchup in bottles labeled as 12 ounces, is (11.94, 11.98), this does *not* tell you that 99% of 12-ounce ketchup bottles contain between 11.94 and 11.98 ounces of ketchup. Nor does it tell you that 99% of samples of the same size would have sample means in this particular range. The confidence interval is an estimate of the *mean* for *all* bottles in the *population* of interest.

Interpreting the Results of Statistical Analyses

Unfortunately, there is no customary way of reporting the estimates of population characteristics in published sources. Possibilities include

confidence interval
estimate ± margin of error
estimate ± standard error

If the population characteristic being estimated is a population mean, then you may also see

sample mean ± sample standard deviation

If the interval reported is described as a confidence interval, a confidence level should accompany it. These intervals can be interpreted just as you have interpreted the other confidence intervals in this chapter. The confidence level specifies the long-run error rate associated with the method used to construct the interval. For example, a 95% confidence level specifies a 5% long-run error rate.

A form particularly common in news articles is estimate ± margin of error. The margin of error reported is usually two times the standard deviation of the estimate. This method of reporting is a little less formal than a confidence interval and, if the sample size is reasonably large, is roughly equivalent to reporting a 95% confidence interval. You can interpret these intervals as you would a confidence interval with approximate confidence level of 95%.

You must use care in interpreting intervals reported in the form of an estimate ± standard error. Recall from Section 9.2 that the general form of a confidence interval is

estimate ± (critical value)(standard deviation of the estimate)

In journal articles, the estimated standard deviation of the estimate is usually referred to as the *standard error*. The critical value in the confidence interval formula was determined by the form of the sampling distribution of the estimate and by the confidence level.

Notice that the reported form, estimate ± standard error, is equivalent to a confidence interval with the critical value set equal to 1. For a statistic whose sampling distribution is approximately normal (such as the sample mean for a large sample or the sample proportion for a large sample), a critical value of 1 corresponds to an approximate confidence level of about 68%. Because a confidence level of 68% is rather low, you may want to use the given information and the confidence interval formula to convert to an interval with a higher confidence level.

When researchers are trying to estimate a population mean, they sometimes report sample mean ± sample standard deviation. Be particularly careful here. To convert this information into a useful interval estimate of the population mean, you must first convert the sample standard deviation to the standard error of the sample mean (by dividing by $\sqrt{n}$) and then use the standard error and an appropriate critical value to construct a confidence interval.

For example, suppose that a random sample of size 100 is used to estimate the population mean. If the sample resulted in a sample mean of 500 and a sample standard deviation of 20, you might find the published results summarized in any of the following ways:

95% confidence interval for the population mean: (496.08, 503.92)
mean ± margin of error: 500 ± 4
mean ± standard error: 500 ± 2
mean ± standard deviation: 500 ± 20

What to Look for in Published Data

Here are some questions to ask when you encounter interval estimates in published reports.

- Is the reported interval a confidence interval, mean ± margin of error, mean ± standard error, or mean ± standard deviation? If the reported interval is not a confidence interval, you may want to construct a confidence interval from the given information.
- What confidence level is associated with the given interval? Is the choice of confidence level reasonable? What does the confidence level say about the long-run error rate of the method used to construct the interval?
- Is the reported interval relatively narrow or relatively wide? Has the population characteristic been estimated precisely?

For example, the article **"Use of a Cast Compared with a Functional Ankle Brace After Operative Treatment of an Ankle Fracture" (*Journal of Bone and Joint Surgery* [2003]: 205–211)** compared two different methods of immobilizing an ankle after surgery to repair damage from a fracture. The article includes the following statement:

> The mean duration (and standard deviation) between the operation and return to work was 63±13 days (median, sixty-three days; range, thirty three to ninety-eight days) for the cast group and 65±19 days (median, sixty-two days; range, eight to 131 days) for the brace group; the difference was not significant.

This is an example of a case where you must be careful—the reported intervals are of the form estimate ± standard deviation. You can use this information to construct a confidence interval for the mean time between surgery and return to work for each method of immobilization. One hundred patients participated in the study, with 50 wearing a cast after surgery and 50 wearing an ankle brace (random assignment was used to assign patients to treatment groups). Because the sample sizes are both large, you can use the t confidence interval formula

$$\text{mean} \pm (t \text{ critical value})\left(\frac{s}{\sqrt{n}}\right)$$

Copyright 2025 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s). Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

Each sample has df = 50 − 1 = 49. The closest df value in Appendix Table 3 is for df = 40, and the corresponding t critical value for a 95% confidence level is 2.02. The corresponding intervals are

$$\text{Cast: } 63 \pm 2.02\left(\frac{13}{\sqrt{50}}\right) = 63 \pm 3.71 = (59.29, 66.71)$$

$$\text{Brace: } 65 \pm 2.02\left(\frac{19}{\sqrt{50}}\right) = 65 \pm 5.43 = (59.57, 70.43)$$

The chosen confidence level of 95% implies that the method used to construct each of the intervals has a 5% long-run error rate. Assuming that it is reasonable to view these samples as representative of the patient population, you can interpret these intervals as follows: You can be 95% confident that the mean return-to-work time for those treated with a cast is between 59.29 and 66.71 days. You can be 95% confident that the mean return-to-work time for those treated with an ankle brace is between 59.57 and 70.43 days.

These intervals are relatively wide, indicating that the values of the treatment means have not been estimated as precisely as you might like. This is not surprising, given the sample sizes and the variability in each sample. Note that the two intervals overlap. This supports the statement that the difference between the two immobilization methods was not significant. Formal methods for directly comparing two groups, covered in Chapter 11, could be used to further investigate this issue.

A Word to the Wise: Cautions and Limitations

When working with point and confidence interval estimates, here are a few things you need to keep in mind:

1. In order for an estimate to be useful, you must know something about accuracy. You should beware of point estimates that are not accompanied by a margin of error or some other measure of accuracy.
2. A confidence interval estimate that is wide indicates that you don't have very precise information about the population characteristics being estimated. Don't be fooled by a high confidence level if the resulting interval is wide. High confidence, while desirable, is not the same thing as saying you have precise information about the value of a population characteristic.

   The width of a confidence interval is affected by the confidence level, the sample size, and the standard deviation of the statistic used to construct the interval. The best strategy for decreasing the width of a confidence interval is to take a larger sample. It is far better to think about this before collecting data and to use the required sample size formulas to determine a sample size that will result in a confidence interval estimate that is narrow enough to provide useful information.
3. The accuracy of estimates depends on the sample size, not the population size. This may be counter to intuition, but as long as the sample size is small relative to the population size (n less than 10% of the population size), the margin of error for estimating a population proportion with 95% confidence is approximately $2\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$ and for estimating a population mean with 95% confidence is approximately $2\frac{s}{\sqrt{n}}$.

   Notice that each of these involves the sample size n, and both margins of error decrease as the sample size increases. Neither approximate margin of error depends on the population size.

   The size of the population does need to be considered if sampling is without replacement and the sample size is more than 10% of the population size. In this case, a **finite population correction factor** $\sqrt{\frac{N-n}{N-1}}$ is used to adjust the margin of error (the given margin is multiplied by the correction factor). Since this correction factor is always less than 1, the adjusted margin of error is smaller.

4. Assumptions and "plausibility" conditions are important. The confidence interval procedures of this chapter require certain assumptions. If these assumptions are met, the confidence intervals provide a method for using sample data to estimate population characteristics with confidence. When the assumptions associated with a confidence interval procedure are in fact true, the confidence level specifies a correct success rate for the method. However, assumptions (such as the assumption of a normal population distribution) are rarely exactly met in practice. Fortunately, in most cases, as long as the assumptions are approximately met, the confidence interval procedures still work well.

   In general, you can only determine if assumptions are "plausible" or approximately met, and that you are in the situation where you expect the procedure to work reasonably well. This is usually confirmed by knowledge of the data collection process and by using the sample data to check certain "plausibility conditions."

   The formal assumptions for the large-sample z confidence interval for a population proportion are

   a. The sample is a random sample from the population of interest.
   b. The sample size is large enough for the sampling distribution of $\hat{p}$ to be approximately normal.
   c. Sampling is without replacement.

   Whether the random sample assumption is plausible will depend on how the sample was selected and the intended population. Plausibility conditions for the other two assumptions are the following:

   $n\hat{p} \geq 10$ and $n(1 - \hat{p}) \geq 10$ (so the sampling distribution of $\hat{p}$ is approximately normal), and

   n is less than 10% of the population size (so that the formula for the standard deviation of $\hat{p}$ provides a good approximation to the actual standard deviation).

   The formal assumptions for the t confidence interval for a population mean are

   a. The sample is a random sample from the population of interest.
   b. The population distribution is normal, so that the distribution of $t = \dfrac{\bar{x} - \mu}{s/\sqrt{n}}$ has a t distribution.

   As was the case for proportions, the plausibility of the random sample assumption will depend on how the sample was selected and the population of interest. The plausibility conditions for the normal population distribution assumption are the following:

   > A normal probability plot of the data is reasonably straight (indicating that the population distribution is approximately normal)

   or

   > The data distribution is approximately symmetric and there are no outliers. This could be confirmed by looking at a dotplot, boxplot, stem-and-leaf display, or histogram of the data.

   Alternatively, if n is large ($n \geq 30$), the sampling distribution of $\bar{x}$ will be approximately normal even for nonnormal population distributions. This implies that use of the t interval is appropriate even if population normality is not plausible.

   In the end, you must decide that the assumptions are met or that they are "plausible" and that the inferential method used will provide reasonable results. This is also true for the inferential methods introduced in the chapters that follow.

5. Watch out for the "±" when reading published reports. Don't fall into the trap of thinking confidence interval every time you see a ± in an expression. As was discussed earlier in this section, published reports are not consistent, and in addition to confidence intervals, it is common to see estimate ± standard error and estimate ± sample standard deviation reported.

Exercises 9.55–9.57

9.55 The following quote is from the article **"Here's How Different Generations in the US Spend Their Income" (www.weforum.org, October 5, 2022)**:

> Generation X—those born from 1965 to 1980—spent the most money last year, with average expenditures of $83,357.

What additional information would you want in order to evaluate the accuracy of this estimate? Explain.

9.56 Authors of the news release titled **"Major Gaps Still Exist Between the Perception and the Reality of Americans' Internet Security Protections, Study Finds" (The National Cyber Security Alliance)** estimated the proportion of Americans who claim to have a firewall installed on their computer to protect them from computer hackers to be 0.80 based on a survey conducted by the Zogby market research firm. They also estimated the proportion of those who actually have a firewall installed to be 0.42, based on checkups performed by Norton's PC Help software. The following quote is from the news release:

> For the study, NCSA commissioned a Zogby survey of more than 3000 Americans and Symantec conducted checkups of 400 Americans' personal computers performed by PC Help by Norton (www.norton.com/tuneup). The Zogby poll has a margin of error of +/− 1.6% and the checkup has a margin of error of +/− 5%.

Explain why the margins of error for the two estimated proportions are different.

9.57 The paper **"The Curious Promiscuity of Queen Honey Bees (*Apis mellifera*): Evolutionary and Behavioral Mechanisms" (*Annals of Zoology Fennici* [2001]: 255–265)** describes a study of the mating behavior of queen honeybees. The following quote is from the paper:

> Queens flew for an average of 24.2 ± 9.21 minutes on their mating flights, which is consistent with previous findings. On those flights, queens effectively mated with 4.6 ± 3.47 males (mean ± SD).

a. The intervals reported in the quote from the paper were based on data from the mating flights of $n = 30$ queen honeybees. One of the two intervals reported is stated to be a confidence interval for a population mean. Which interval is this? Justify your choice.

b. Use the given information to construct a 95% confidence interval for the mean number of partners on a mating flight for queen honeybees. For purposes of this exercise, assume that it is reasonable to consider these 30 queen honeybees as representative of the population of queen honeybees.

Section 9.5 Bootstrap Confidence Intervals for a Population Proportion (Optional)

In Section 9.2, data from a large random sample were used to construct a confidence interval for a population proportion, p. The large-sample confidence interval has the form

$$\hat{p} \pm (z \text{ critical value})\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

To use this confidence interval formula, you need to know that the sample is a random sample from the population of interest (or is selected in a way that makes it reasonable to think that the sample is representative of the population). In addition, you need to know that the sampling distribution of $\hat{p}$ is approximately normal. This is reasonable when the sample size is large $(n\hat{p} \geq 10$ and $n(1-\hat{p}) \geq 10)$ but isn't necessarily the case when the sample size is small. This section introduces an alternative method that can be used to obtain a confidence interval for a population proportion that doesn't require a large sample size.

A Bootstrap Confidence Interval for a Population Proportion

When a sample proportion $\hat{p}$ is used to estimate a population proportion, the value of $\hat{p}$ is not likely to be exactly equal to the value of the population proportion. But if the sample is selected in a reasonable way, you expect that the value of $\hat{p}$ will be somewhere around the value of the population proportion. A confidence interval quantifies what is meant by "around the value of the population proportion." When the assumptions of the large-sample confidence interval are reasonable, the confidence interval based on a sampling distribution of $\hat{p}$ that is approximately normal can be used to determine the largest value that the difference between the observed sample proportion and the actual value of the population proportion is likely to be for a specific confidence level.

However, sometimes the distribution of $\hat{p}$ is not approximately normal. In this case, constructing a confidence interval still requires knowing something about how far away the sample proportion is likely to be from the value of the population proportion. For example, suppose that the sample proportion is $\hat{p} = 0.32$ and that it is unlikely to be smaller than the actual value of the population proportion by more than 0.04. Then you would think that the population proportion is less than 0.36 (from $0.32 + 0.04$). If it is also unlikely that the sample proportion is greater than the population proportion by more than 0.05, then you could say that you think that the population proportion is greater than 0.27 (from $0.32 - 0.05$). A reasonable interval estimate of the population proportion based on the sample would then be (0.27, 0.36). **Bootstrapping** is a way to determine what number to add and what number to subtract from the sample proportion to form a confidence interval.

Bootstrapping works by considering how far sample proportions from random samples of size n tend to be from the value of the population proportion if the value of the population proportion were known. Taking many different random samples of size n from such a population and calculating the sample proportion for each one creates a distribution of these sample proportions. Looking at how the sample proportions cluster around the value of the population proportion provides information about variability. But of course, in practice you can't do this because the value of the population proportion is unknown.

Instead, consider a hypothetical population that you expect to be very similar to the population that the sample is actually from. Examining sample proportions from this hypothetical population provides information about variability in $\hat{p}$ values. To do this, bootstrapping uses the observed sample proportion as the proportion for the hypothetical population.

> To create a bootstrap confidence interval, many random samples of size n are taken from a hypothetical population to form a **bootstrap distribution**. The variability in this distribution indicates how far $\hat{p}$ values for samples from this hypothetical population might be from the original observed value of $\hat{p}$. Knowing how far these simulated $\hat{p}$ values tend to fall from the observed value of $\hat{p}$ indicates how far the observed value of $\hat{p}$ is likely to be from the value of p in the actual population.

For a 95% confidence level, using the boundaries that capture the middle 95% of the simulated bootstrap distribution is equivalent to determining the endpoints of a confidence interval, which is represented as $\hat{p}$ minus a number and $\hat{p}$ plus a number. This interval is called a **bootstrap confidence interval** estimate for p, and it is based on simulation rather than on knowing that the sampling distribution of $\hat{p}$ is approximately normal.

> For bootstrap confidence intervals, the number subtracted from $\hat{p}$ and the number added to $\hat{p}$ won't always be equal since the bootstrap distribution may not be symmetric.

Taking a random sample from a hypothetical population that has a proportion of successes equal to the original sample proportion is equivalent to selecting a random sample *with replacement* from the original sample. The process of drawing new samples from an original representative sample is called **resampling**, because new samples are taken directly from the original sample that was collected. This method is also called "bootstrapping" because it is like "pulling yourself up by the bootstraps," in the sense that nothing more than the original data collected in one sample is used to generate information about sample-to-sample variability in the sample proportion, $\hat{p}$.

Example 9.11 Generating a Bootstrap Distribution

Suppose that a random sample of 10 students at a school is selected, and each student is asked if they spend more than 3 hours a day on social media. The sample will be used to estimate the proportion of students at the school who would respond "Yes" to this question. If the responses for the people in the sample were those shown below, the value of the sample proportion of "Yes" responses would be $\hat{p} = \frac{5}{10} = 0.5$.

Student Number	1	2	3	4	5	6	7	8	9	10
Response	No	Yes	Yes	No	No	Yes	No	No	Yes	Yes

To create a bootstrap distribution, consider a hypothetical population in which the proportion of successes is 0.5, and then take samples of size 10 from this hypothetical population. One way to do this is to select a random sample of size 10 *with replacement* from the original sample using random numbers. Ten random numbers, selected with replacement from the numbers from 1 to 10, are shown below. Also shown are the responses from the original sample associated with each of these random numbers.

Random Number	10	9	3	8	3	2	8	3	6	9
Response	Yes	Yes	Yes	No	Yes	Yes	No	Yes	Yes	Yes

The proportion of "Yes" responses in this simulated sample is $\hat{p} = \frac{8}{10} = 0.8$, even though the proportion in the original sample was 0.5.

To form a bootstrap distribution, you resample from the original sample many times. These simulated samples provide information about the sample-to-sample variability in the sample proportions, $\hat{p}$, and the bootstrap distribution can then be used to determine an interval of plausible values for p associated with a specified confidence level.

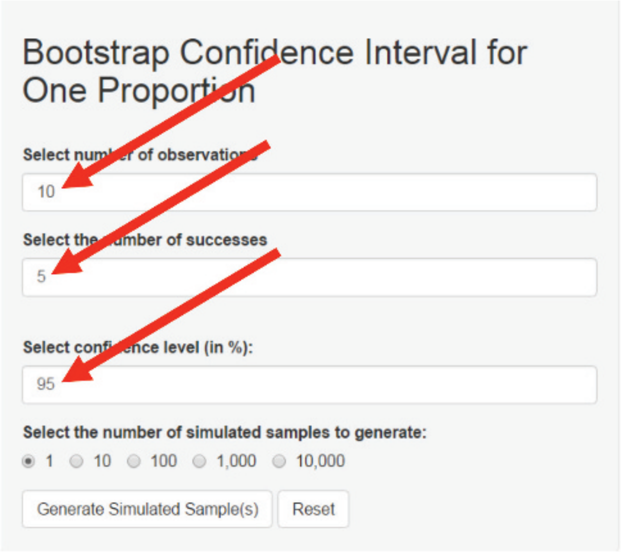
Bootstrap Confidence Intervals for One Proportion

This section explains how to generate a bootstrap distribution and construct a bootstrap confidence interval for a population proportion using one of the Shiny web apps that accompany this text. These web apps are located at **www.cengage.com/Peck.html**. There are also a number of other resources for constructing bootstrap confidence intervals (for example, see the StatKey apps at **lock5stat.com/StatKey/**).

Example 9.12 Generating a Bootstrap Confidence Interval for a Population Proportion (Example 9.11 continued)

Recall that in the previous example, a random sample of 10 students produced a sample proportion who spend more than 3 hours a day on social media of $\hat{p} = 0.50$. To use the Shiny app to construct a bootstrap confidence interval for the proportion of students at the school who spend more than 3 hours a day on social media, follow the instructions below.

Enter the number of observations and the number of successes into the Shiny app titled "Bootstrap Confidence Interval for One Proportion." In this example, the number of observations, also known as the sample size, is $n = 10$, and five successes were observed in the original sample. The confidence level in this example will be 95%, so this value should be entered as well, or recognize that 95% is the default.



Bootstrap Confidence Interval for One Proportion

Select number of observations: 10

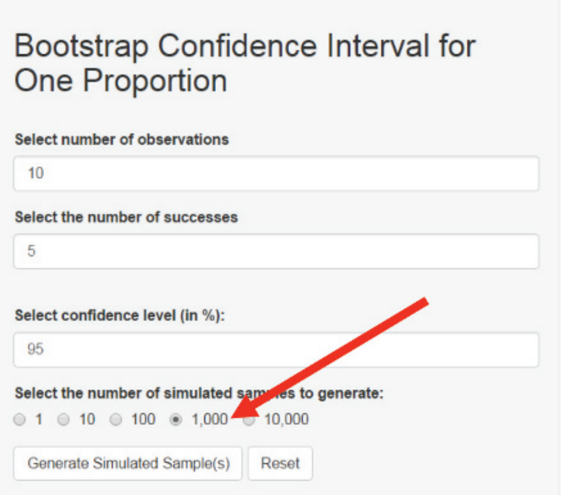
Select the number of successes: 5

Select confidence level (in %): 95

Select the number of simulated samples to generate:

● 1 ○ 10 ○ 100 ○ 1,000 ○ 10,000

Select the number of simulated bootstrap samples to be generated. In this example, choose 1000 simulated samples. Then, click “Generate Simulated Sample(s).”



Bootstrap Confidence Interval for One Proportion

Select number of observations: 10

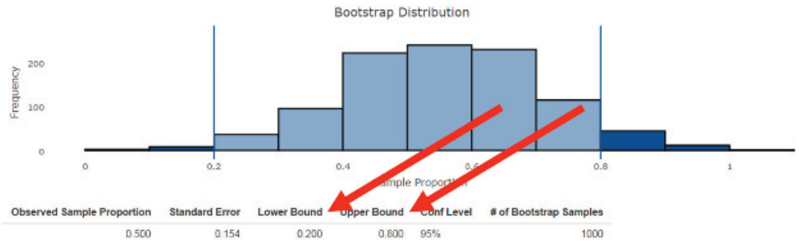
Select the number of successes: 5

Select confidence level (in %): 95

Select the number of simulated samples to generate:

○ 1 ○ 10 ○ 100 ● 1,000 ○ 10,000

The Shiny app identifies the “Lower Bound” to be the value of the sample proportion that has 2.5% of the simulated proportions below and the “Upper Bound” to be the value that has 2.5% of the simulated proportions above:



Observed Sample Proportion	Standard Error	Lower Bound	Upper Bound	Conf Level	# of Bootstrap Samples
0.500	0.154	0.200	0.800	95%	1000

Different sets of simulated samples may produce slightly different results. Based on this output from the Shiny app, 2.5% of the simulated sample proportions fall at or below 0.2, and 2.5% of the simulated sample proportions fall at or above 0.8. A 95% bootstrap

Copyright 2025 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s). Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

confidence interval for the proportion of students at the school who would respond that they spend more than 3 hours a day on social media is (0.2, 0.8). You can be 95% confident that the actual proportion of "Yes" responses falls between 0.2 and 0.8. Notice that this interval is very wide, with plausible values for the population proportion ranging from 0.2 all the way to 0.8. This is a function of the very small sample size. It is difficult to estimate population proportions accurately with a small sample, no matter what method is used!

Because a bootstrap confidence interval is based on a distribution of simulated proportions, the interval that is generated may vary from one simulation to another. But if the confidence interval is based on many simulated proportions, different bootstrap confidence intervals based on the same sample won't differ substantially from one another.

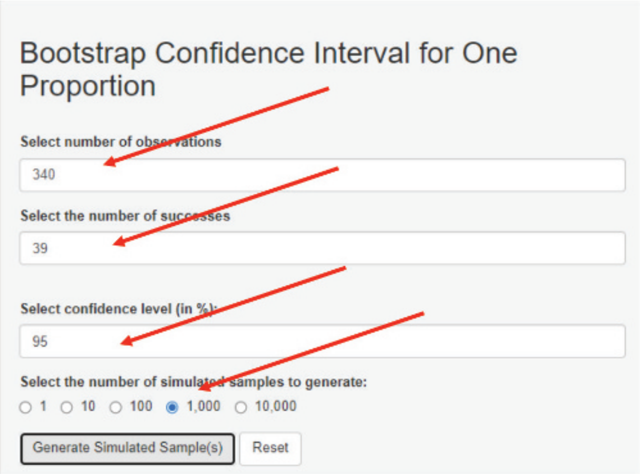
Next, consider a more realistic example.

Example 9.13 Shark Bites Revisited

Recall that in Example 9.4 the large-sample confidence interval formula was used find a 95% confidence interval for p, the population proportion of adult residents of North Carolina who believe that sharks are to blame when sharks bite people. The observed value of the sample proportion was $\hat{p} = \dfrac{39}{340} = 0.115$, and the resulting confidence interval was (0.082, 0.148).

You can use the bootstrap method as an alternative way to find a 95% confidence interval for p.

In the Shiny app "Bootstrap Confidence Interval for One Proportion," enter the sample size and the number of successes. In Example 9.4, the sample size is $n = 340$, and 39 successes were observed in the original sample. Select a 95% confidence level and then choose 1000 simulated samples. Click on "Generate Simulated Sample(s)."



Bootstrap Confidence Interval for One Proportion

Select number of observations

340

Select the number of successes

39

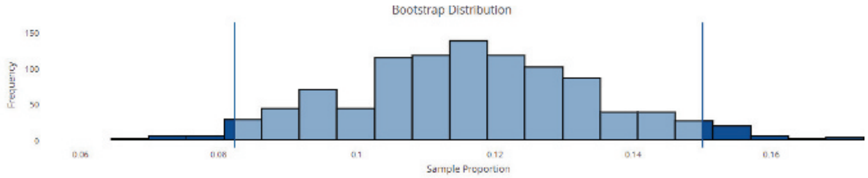
Select confidence level (in %):

95

Select the number of simulated samples to generate:

○ 1 ○ 10 ○ 100 ● 1,000 ○ 10,000

For the simulation shown here, the Shiny app identifies the value that has 2.5% of the simulated proportions below and the value that has 2.5% of the simulated proportions above:



Observed Sample Proportion	Standard Error	Lower Bound	Upper Bound	Conf Level	# of Bootstrap Samples
0.115	0.018	0.082	0.150	95%	1000

The Shiny app reports that 2.5% of the simulated sample proportions fall at or below 0.082, and that 2.5% of the simulated sample proportions fall at or above 0.150. A bootstrap confidence interval for the actual proportion of adult residents of North Carolina who believe

Copyright 2025 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s). Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

that sharks are to blame when sharks bite people is (0.082, 0.150). Assuming that the sample is representative of the population, you can be 95% confident that the actual proportion of adult residents of North Carolina who believe that sharks are to blame when sharks bite people falls between 0.082 and 0.150 (between 8.2% and 15.0%).

Notice that this bootstrap confidence interval is very close to the interval provided by the large-sample method in Section 9.2, which is not surprising since the conditions for the large-sample confidence interval are satisfied.

What Happens When the Conditions for the Large-Sample Interval Are Not Met?

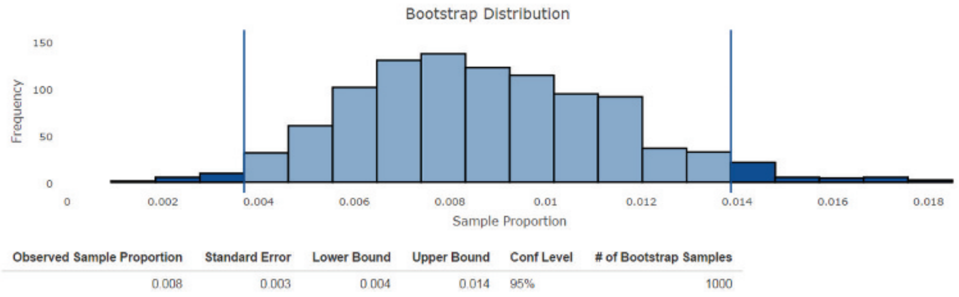
In the previous example, the bootstrap confidence interval turned out to be nearly the same as the large-sample confidence interval. It is not surprising to find that the bootstrap confidence interval and the large-sample confidence interval are similar when the sample size is large. In this case, both methods provide an appropriate way to obtain a confidence interval. But the bootstrap method can still be used even if the sample size is not large enough to satisfy the conditions for the large-sample confidence interval. This is illustrated in the following example.

Example 9.14 Liver Injuries in Newborns

The article **"Severe Liver Injury While Using Umbilical Venous Catheter: Case Series and Literature Review" (*American Journal of Perinatology* [2014]: 965–974)** describes a study of newborns who were placed in intensive care and required insertion of an umbilical vein catheter so that fluids could be administered. Researchers found that 9 out of the 1081 newborns studied developed catheter-associated liver injury. The authors were interested in estimating the proportion of newborns who suffer liver injury as a result of the use of umbilical vein catheters.

The researchers considered this sample of 1081 infants to be representative of the population of newborns who require use of the catheter. The sample proportion with liver injury is $\hat{p} = \dfrac{9}{1081} = 0.00833$. Notice that although the sample size is 1081, because $n\hat{p} = (1081)(0.00833) = 9$, the sample size is not large enough to justify the use of the large-sample confidence interval for a population proportion.

The Shiny app "Bootstrap Confidence Interval for One Proportion" was used to generate a bootstrap distribution based on 1000 simulated sample proportions.



Observed Sample Proportion	Standard Error	Lower Bound	Upper Bound	Conf Level	# of Bootstrap Samples
0.008	0.003	0.004	0.014	95%	1000

A 95% bootstrap confidence interval for the population proportion, p, is (0.004, 0.014). Based on this sample, you can be 95% confident that the actual proportion of newborns with umbilical catheters who suffer liver injury is somewhere between 0.004 and 0.014. Because the conditions for the large-sample interval are not met, the bootstrap interval is a more appropriate choice.

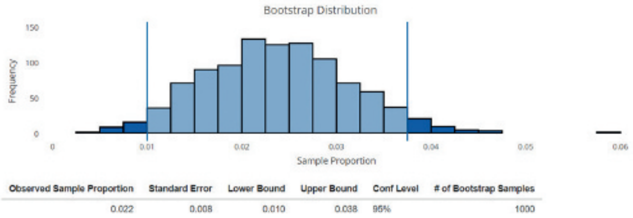
Copyright 2025 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s). Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

Exercises 9.58–9.65

9.58 Data from a survey of full-time and part-time workers are summarized in the article **"This Is What People Hate Most About Their Boss, Survey Shows" (BestLife, August 20, 2021)**. The percentage of survey recipients who identified being ungrateful as the thing they liked least about their boss was 2.25%. Suppose that this percentage was based on a representative sample of 400 workers, with 9 choosing ungrateful.

a. Would it be appropriate to use the large-sample confidence interval for a population proportion to estimate the proportion of full-time and part-time workers who would identify being ungrateful as the thing they liked least about their boss? Justify your answer.

b. Use the accompanying output from the "Bootstrap Confidence Interval for One Proportion" Shiny app to report a 95% confidence interval for the proportion who would choose ungrateful.



Observed Sample Proportion	Standard Error	Lower Bound	Upper Bound	Conf Level	# of Bootstrap Samples
0.022	0.008	0.010	0.038	95%	1000

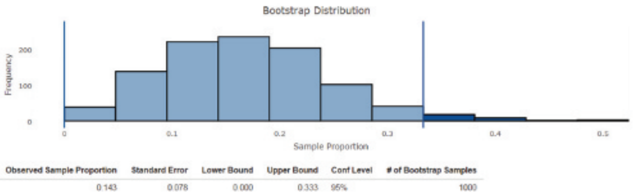
9.59 The article **"Report: More Than Half of DC-Area Millennials Are Using Ride-Hailing Apps" (washingtonian.com)** refers to a study summarized at the following website: **wbaresearch.com/wp-content/uploads/2016/06/Transportation-MarkeTrak-Spring-20161.pdf**. The study indicates that 21% of Washington-area adults who are 55 years old and older have used transportation apps such as Uber or Lyft at least once.

Suppose that a small local transportation service for older residents is monitoring usage of app-based transportation. The service conducted a survey of a random sample of 21 of its regular customers who are 55 or older and found that 3 of them had tried Uber or Lyft at least once.

a. Would it be appropriate to use the large-sample confidence interval for a population proportion to estimate the proportion of the transportation service customers who have tried Uber or Lyft at least once? Explain.

b. Would it be appropriate to use a bootstrap confidence interval for a population proportion to estimate the proportion of the transportation service customers who have tried Uber or Lyft at least once? Explain.

c. Use the accompanying output from the Shiny app to report a 95% bootstrap confidence interval for the population proportion of customers 55 or older who have used Uber or Lyft at least once. Interpret the confidence interval in context.



Observed Sample Proportion	Standard Error	Lower Bound	Upper Bound	Conf Level	# of Bootstrap Samples
0.143	0.078	0.000	0.333	95%	1000

d. Is the value obtained in the study for Washington-area adults who are over 55 years old who have used Uber or Lyft at least once (21%) in the bootstrap confidence interval? What does this indicate?

9.60 An article titled **"The Latest on Workplace Monitoring and Surveillance" (American Management Association, November 17, 2014)** referred to the "2007 Electronic Monitoring & Surveillance Survey." In a summary of survey results submitted by 304 U.S. businesses, 85 of these businesses had fired workers for e-mail misuse.

Suppose that it is reasonable to regard these 304 businesses as a representative sample of businesses in the United States. Use the "Bootstrap Confidence Interval for One Proportion" Shiny app to generate a bootstrap confidence interval for the proportion of U.S. businesses who have fired workers for e-mail misuse. Interpret the interval in context.

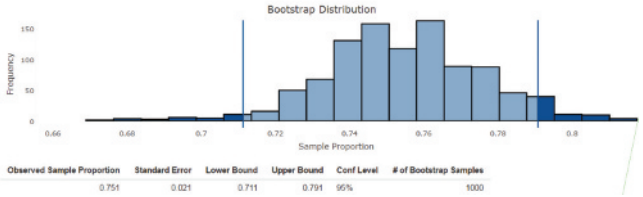
9.61 In a study of long-term effects of infection from COVID-19, doctors in Canada followed up with 47 patients within 3 months after recovery **(*Journal of the American Medical Association Cardiology* Online, January 12, 2022)**. The authors found that 8 patients had signs of myocardial inflammation (an inflammation of the heart wall).

a. Explain why it would not be appropriate to use a large-sample confidence interval for a population proportion to estimate the proportion of COVID-19 patients who experience myocardial inflammation.

b. Assuming that the sample is representative of patients who have had a COVID-19 infection, use the "Bootstrap Confidence Interval for One Proportion" Shiny app to report a 90% confidence interval for the proportion of COVID-19 patients who experience myocardial inflammation.

9.62 A survey of a representative sample of 478 U.S. employers determined that 359 ranked stress as their top health and productivity concern **(globenewswire.com, June 29, 2016)**.

a. Use the accompanying output from the "Bootstrap Confidence Interval for One Proportion" Shiny app to report a 95% bootstrap confidence interval for the proportion of all U.S. employers who would rank stress as their top health and productivity concern. Interpret the confidence interval in context.



Observed Sample Proportion	Standard Error	Lower Bound	Upper Bound	Conf Level	# of Bootstrap Samples
0.751	0.021	0.711	0.791	95%	1000

b. A number of international employers were also surveyed. If the international employers had a similar rate of identifying stress as their top health and productivity concern, and if the results from international employers were included in the sample, would the width of the resulting confidence interval remain the same, decrease, or increase? Explain.

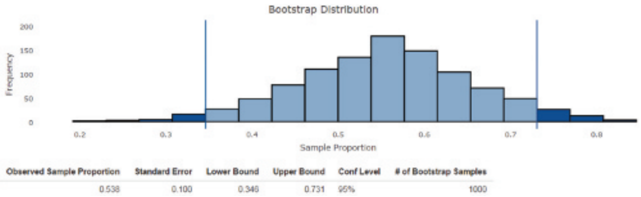
9.63 In mid-2016 the United Kingdom (UK) voted to withdraw from the European Union (an event known as "Brexit"), causing economic concerns throughout the world. One indicator that economists use to monitor the health of the economy is the proportion of residential properties offered for sale at auction that are successfully sold.

An article titled **"Going, Going, Gone through the Roof—Sky's the Limit at Auction" (estateagenttoday.co.uk, October 22, 2016)** reported the success rate for a sample of 26 residential properties offered for sale at auctions in the UK in the summer of 2016. For this sample of properties, 14 of the 26 residential properties were successfully sold. Suppose it is reasonable to consider these 26 properties as representative of residential properties offered at auction in the post-Brexit UK.

a. Would it be appropriate to use the large-sample confidence interval for a population proportion to estimate the proportion of residential properties successfully sold at auction in the post-Brexit UK? Explain.

b. Would it be appropriate to use a bootstrap confidence interval for a population proportion to estimate the proportion of residential properties successfully sold at auction in the post-Brexit UK? Explain.

c. Use the accompanying output from the "Bootstrap Confidence Interval for One Proportion" Shiny app to report a 95% bootstrap confidence interval for the population proportion of residential properties successfully sold at auction in the post-Brexit UK. Interpret the confidence interval in context.



Observed Sample Proportion	Standard Error	Lower Bound	Upper Bound	Conf Level	# of Bootstrap Samples
0.538	0.100	0.346	0.731	95%	1000

d. The success rate for properties sold at auction throughout the UK during a period of time one year earlier—in 2015—was 72%. Does this value fall within the bootstrap confidence interval reported in Part (c)? What does this indicate?

9.64 The report **"One in Three American Households Are Stuck in a Relationship with a Financial Services Provider They Don't Trust" (businesswire.com, June 29, 2016)** estimated that 31% of American households feel obliged to do business with one or more financial services companies they distrust. This estimate is based on a representative sample of 1056 households.

Use the "Bootstrap Confidence Interval for One Proportion" Shiny app to generate a 95% bootstrap confidence interval for the proportion of all U.S. households that feel obliged to do business with one or more financial services companies they distrust. Interpret the interval in context.

9.65 A 2016 study of 120 U.S. brand-name products found that 70% were active on Snapchat **(businessinsider.com, June 15, 2016)**. The researchers conducting the study used bootstrap methods to determine a confidence interval.

Suppose that it is reasonable to consider this sample of brand-name products as representative of all brand-name products. Use the "Bootstrap Confidence Interval for One Proportion" Shiny app to find a 95% confidence interval for the proportion of all brand-name products that are active on Snapchat and interpret the interval in context.

Section 9.6 Bootstrap Confidence Intervals for a Population Mean (Optional)

The sample mean $\bar{x}$ is an estimate of the population mean μ, but you wouldn't expect that the value of $\bar{x}$ will be exactly equal to the population mean. However, for representative samples, you would expect that the value of $\bar{x}$ will be near the value of μ. The notion of being near the value of a population mean can be represented by a confidence interval.

Copyright 2025 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s). Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

In Section 9.3, data from a sample were used to construct a confidence interval for a population mean, μ. The confidence interval introduced there has the form

$$\bar{x} \pm (t \text{ critical value}) \frac{s}{\sqrt{n}}$$

To use this confidence interval, you need to know that the population distribution is approximately normal or the sample size is large. If you are not certain about whether the assumptions needed for a large-sample confidence interval for μ are reasonable, a simulation-based approach called **bootstrapping** can be used.

Consider a hypothetical population that is similar to the population that the observed sample is actually from. By examining the distribution of sample means for samples taken from this hypothetical population, it is possible to use bootstrapping to find a confidence interval for the population mean. Bootstrapping uses random samples (called bootstrap samples) from a hypothetical population represented by the data in the observed sample, which is thought to be representative of the population. The distribution of sample means from the bootstrap samples is called a **bootstrap distribution**. The variability in this distribution indicates how far $\bar{x}$ values for samples from this hypothetical population might be from the original observed value of $\bar{x}$. Knowing how far simulated $\bar{x}$ values tend to fall from the observed value of $\bar{x}$ provides information about how far the observed value of $\bar{x}$ is likely to be from the value of μ in the actual population.

For a 95% confidence level, using the boundaries that capture the middle 95% of the bootstrap distribution is equivalent to getting a value for $\bar{x}$ minus a number, and for $\bar{x}$ plus a number (although the numbers subtracted and added won't always be the same since the bootstrap distribution may not be symmetric). This interval is a **bootstrap confidence interval** estimate for μ, and it is based on simulation rather than on knowing that the sampling distribution $\bar{x}$ is at least approximately normal.

Example 9.15 Electric Car Energy Consumption (Revisited)

● Example 9.9 of Section 9.3, gave data on "miles per gallon-equivalent" (MPG-e), a measure of how far an electric vehicle can travel on 33.7 kWh of electricity, which is approximately equivalent to 1 gallon of gas. For example, a rating of 40 MPG-e for an electric car is comparable to a gas-powered car that gets 40 miles per gallon of gas.

Consider the data ➤ The MPG-e rating data for a random sample of seven 2021 model electric cars is displayed in the table below. The mean of this sample is $\bar{x} = 102.86$ MPG-e.

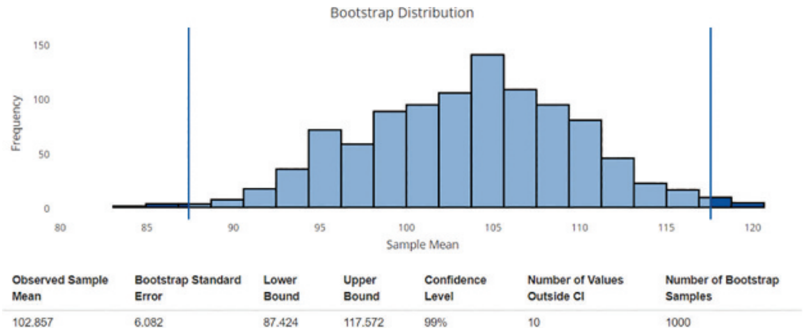
Car ID	1	2	3	4	5	6	7
MPG-e	75	92	96	113	129	111	104

Do the work ➤ You begin by resampling from this original sample. That is, you generate bootstrap samples by selecting at random *with replacement* from the original sample. Here is one bootstrap sample from the original sample:

Resampled Car ID	3	2	1	4	3	5	1
MPG-e	96	92	75	113	96	129	75

The sample mean for this bootstrap sample is $\bar{x} = 96.57$. You can repeat this process many times, and the resulting bootstrap distribution of $\bar{x}$ values provides information about sampling variability that can be used to find confidence intervals. Output on the next page from the Shiny app titled "Bootstrap Confidence Interval for One Mean," can be used to obtain a 99% confidence interval for μ. This Shiny app can be found in the app collection at http://www.cengage.com/Peck.

● **Data set available online**



Observed Sample Mean	Bootstrap Standard Error	Lower Bound	Upper Bound	Confidence Level	Number of Values Outside CI	Number of Bootstrap Samples
102.857	6.082	87.424	117.572	99%	10	1000

Interpret the results ➤ The bootstrap approach produces a simulation-based bootstrap confidence interval for the population mean, μ, from 87.42 to 117.57 MPG-e. You can be 99% confident that this interval contains the population mean. This interval represents a set of plausible values for the population mean, μ.

The one-sample t confidence interval calculated in Example 9.9 was (78.60, 127.12). Notice that the bootstrap confidence interval is narrower than the one-sample t interval.

Exercises 9.66–9.71

● **Data set available online**

9.66 ● *Consumer Reports* published the following estimated annual fuel cost (in dollars) for a sample of electric 2021 car models **(www.consumerreports.org)**.

Make and Model	Annual Fuel Cost
Tesla Model S Performance	700
BMW I3	600
Ford Mustang Mach E	800
Nissan Leaf	600
Tesla Model X Performance	650
Porsche Taycan	900
Volkswagen ID.4	700
MINI Cooper SE	600

Use the output at the bottom of the page from the Shiny App “Bootstrap Confidence Interval for One Mean” to construct and interpret a 95% confidence interval for the population mean annual fuel cost for electric 2021 model cars. You can assume that this sample is representative of the population of 2021 electric car models.

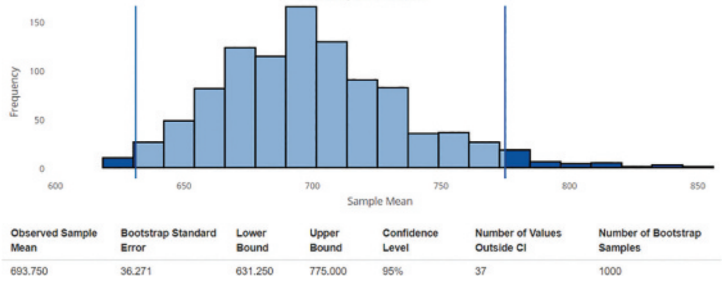
9.67 ● Just how dirty is the money you carry in your wallet? The authors of the article **“Dirty Money: New Research Reveals the Filthiness of Our Cash, Cards, & Coins” (lendedu.com, 2020)** measured the bacteria on the surface of credit and debit cards, bills, and coins. Germ scores (with higher scores corresponding to dirtier surfaces) for 12 coins of various denominations, arranged in order from greatest germ score to lowest, are given below. You may consider these germs scores to be a representative sample of germs scores for coins.

356	329	228	227	109	92
84	60	52	33	32	25

a. Construct a graphical display for the data. What characteristic of the distribution indicates that the methods based on the t distribution may not be appropriate?

b. Use a bootstrap confidence interval (Shiny app: “Bootstrap Confidence Interval for One Mean”) to estimate the population mean germ score for coins using a 95% confidence interval.

Output for Exercise 9.66



Observed Sample Mean	Bootstrap Standard Error	Lower Bound	Upper Bound	Confidence Level	Number of Values Outside CI	Number of Bootstrap Samples
693.750	36.271	631.250	775.000	95%	37	1000

Copyright 2025 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s). Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

9.68 ● *Consumer Reports* provides ratings for refrigerators, including energy cost per year **(consummerreports.org).** Energy cost (in dollars) for a random sample of 11 refrigerator models manufactured in 2021 are given in the accompanying table.

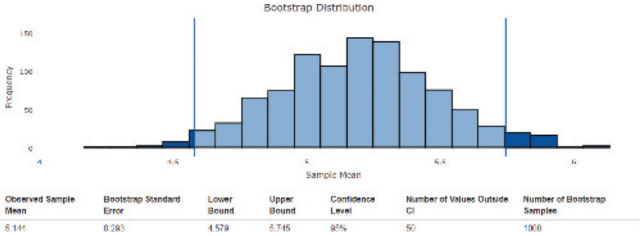
45 52 43 45 48 61 51 53 51 53 54

Use the output at the bottom of the page from the Shiny App "Bootstrap Confidence Interval for One Mean" to construct and interpret a 95% bootstrap confidence interval for the mean annual energy cost for the population of all 2021 model refrigerators.

9.69 ● *Consumer Reports* provides ratings for televisions, including energy cost per year **(consumerreports.org)**. Energy cost data for a sample of 13 small televisions (29-inch and smaller) are displayed in the following table:

Make and Model	Energy Cost (dollars)
Samsung UN28H4000	7.17
LG 28LF4520	5.17
LG 24LF4520	4.47
LG 28LH4530	5.00
Vizio D28h-C1	5.06
Vizio D28h-D1	7.00
LG 22LH4530	4.00
Element ELEFW248	4.00
Vizio D24-D1	6.00
LG 24LH4530	4.00
Insignia NS-24ER310NA17	6.00
Seiki SE23HEB2	4.00
Insignia NS-24D510NA17	5.00

Suppose that this sample is representative of the population of all small televisions. Use the accompanying output from the Shiny app to construct and interpret a 95% bootstrap confidence interval for the mean annual energy cost for the population of all small televisions.



Observed Sample Mean	Bootstrap Standard Error	Lower Bound	Upper Bound	Confidence Level	Number of Values Outside CI	Number of Bootstrap Samples
5.144	0.283	4.579	5.745	95%	50	1000

9.70 ● The *Economist* collects data each year on the price of a Big Mac in various countries around the world. A sample of McDonald's restaurants in Europe in 2022 resulted in the following Big Mac prices (after conversion to U.S. dollars):

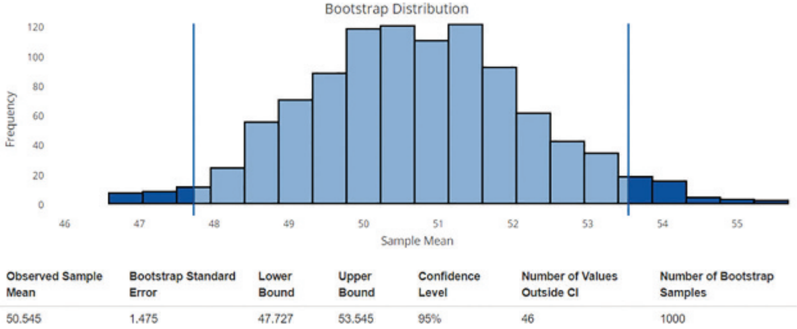
6.98	4.10	4.82	4.95
4.81	6.39	2.40	5.78

a. What characteristic of the sample indicates that the methods based on the *t* distribution may not be appropriate?

b. Use the Shiny app output at the bottom of the page to estimate the 2022 mean price of a Big Mac in Europe using a 95% bootstrap confidence interval.

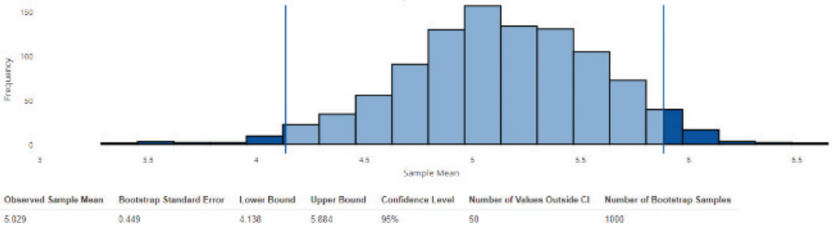
9.71 ● Major League Baseball (MLB) includes two groups of teams, in "leagues." There are 15 teams in each of the American League (AL) and the National League (NL). Since 1997, teams in each of the

Output for Exercise 9.68



Observed Sample Mean	Bootstrap Standard Error	Lower Bound	Upper Bound	Confidence Level	Number of Values Outside CI	Number of Bootstrap Samples
50.545	1.475	47.727	53.545	95%	46	1000

Output for Exercise 9.70



Observed Sample Mean	Bootstrap Standard Error	Lower Bound	Upper Bound	Confidence Level	Number of Values Outside CI	Number of Bootstrap Samples
5.029	0.449	4.138	5.884	95%	50	1000

leagues play teams from the other league in "interleague" regular-season games.

One way to determine whether one league is stronger than the other is to consider the interleague winning percentages for all the teams in one of the two leagues, say, the National League, for one season. For purposes of this exercise, consider the interleague games played in the 2022 season to be a representative sample of the performance of the teams in a population of potential future seasons.

Here are the 2022 interleague winning percentages for the 15 NL teams:

Team	W	L	Winning Percentage
ARI	12	8	60
ATL	13	7	65
CHC	6	14	30
CIN	12	8	60
COL	9	11	45
LAD	15	5	75
MIA	8	12	40
MIL	15	5	75
NYM	9	11	45
PHI	9	11	45
PIT	4	16	20
SD	8	12	40
SF	10	10	50
STL	10	10	50
WSH	8	12	40

(continued)

a. What characteristic of this sample of NL team interleague winning percentages indicates that the methods based on the *t* distribution may not appropriate?

b. Estimate the mean interleague winning percentage for the population of NL teams using a 95% bootstrap confidence interval.

c. Use the results from Part (b) to explain whether it is reasonable to say that the National League or the American League performs significantly better than the other in interleague play.

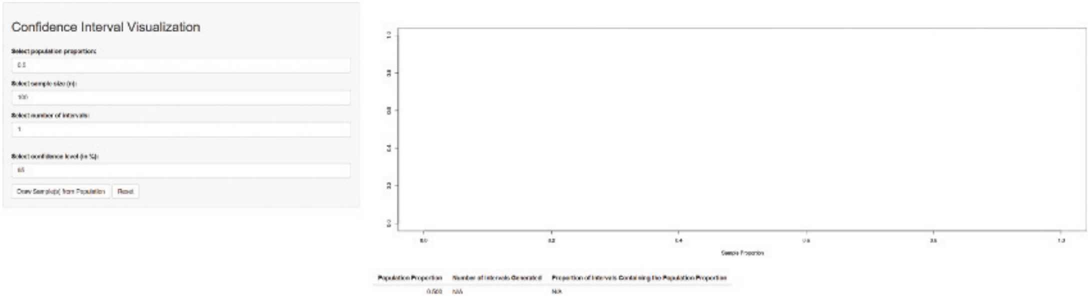
Chapter Activities

Activity 9.1 Getting a Feel for Confidence Level Technology Activity

The purpose of this activity is to explore the meaning of the confidence level associated with a confidence interval.

This activity uses the Shiny app titled CI Proportion Visualization that is part of the collection of Shiny web apps that accompany this text. This app can be found at **www.cengage.com/Peck**.

Open the app. You should see a screen like the one shown here.



Confidence Interval Visualization

Select population proportion: 0.5

Select sample size (n): 100

Select number of intervals: 1

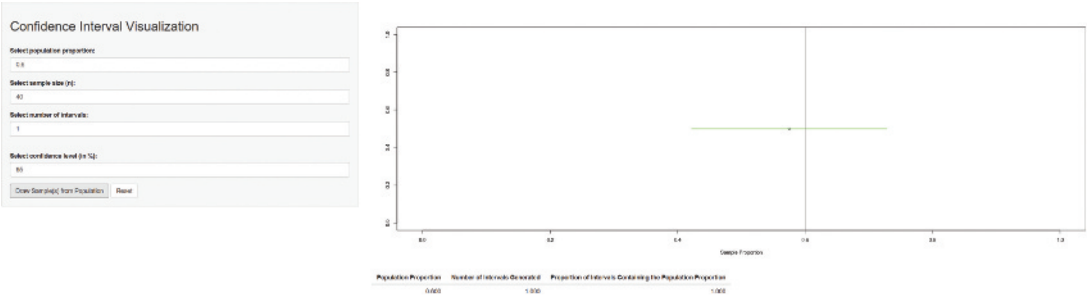
Select confidence level (in %): 95

Population Proportion	Number of Intervals Generated	Proportion of Intervals Containing the Population Proportion
0.500	NA	NA

This app will select a random sample from a population consisting of Successes and Failures. It uses the sample data to construct a confidence interval for the population proportion. This interval is then plotted on the display, and you can see if the resulting confidence interval contains the actual value of the population proportion.

For purposes of this activity, you will take a random sample of size $n = 40$ from a population with a proportion of successes of $p = 0.60$. Change the value in the "Select population proportion" box to 0.60 and change the number in the "Select sample size" box to 40. Leave the number of intervals at 1 and the confidence level at 95. Now click on the "Draw sample(s) from population" button.

You should see a confidence interval appear on the display on the right-hand side. If the interval contains the actual value of the population proportion, the interval is drawn in green. If the value of the population proportion is not in the interval, the interval is shown in red. Your screen should look something like the following:



Confidence Interval Visualization

Select population proportion: 0.6

Select sample size (n): 40

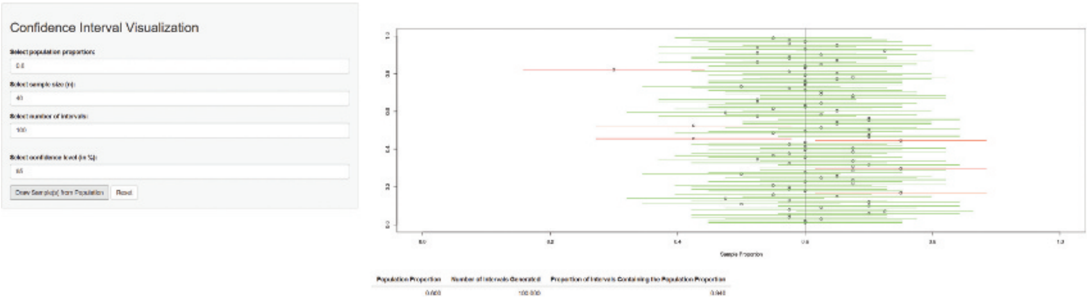
Select number of intervals: 1

Select confidence level (in %): 95

Population Proportion	Number of Intervals Generated	Proportion of Intervals Containing the Population Proportion
0.600	1.000	1.000

Click on the "Draw sample(s) from population" button several more times. Notice how the confidence interval estimate changes from sample to sample. If you were to construct a large number of intervals, the percentage of intervals that contain the value of the population proportion should approximate the capture rate for the confidence interval method.

To look at more than one interval at a time, click on the reset button. Then change the number of intervals to 100. When you click on the "Draw sample(s) from population" button, you should see a screen similar to the one below.



Confidence Interval Visualization

Select population proportion: 0.6

Select sample size (n): 40

Select number of intervals: 100

Select confidence level (in %): 95

Population Proportion	Number of Intervals Generated	Proportion of Intervals Containing the Population Proportion
0.600	100.000	[illegible]

Notice that the table below the display includes the proportion of the intervals generated that include the value of the population proportion.

Next, click on the reset button and change the number of intervals to 1000. Click on the "Draw sample(s) from population" button to see the result of generating 1000 different 95% confidence intervals.

1. How does the proportion of intervals constructed that include $p = 0.60$, the value of the population proportion, compare to 0.95?
2. Experiment with several other confidence levels of your choice, and then answer the following question: In general, is the proportion of confidence intervals that contain $p = 0.60$ close to the stated confidence level?

Activity 9.2 An Alternative Confidence Interval for a Population Proportion Technology Activity

Section 9.2 introduced an alternative confidence interval that can be used in place of the large-sample confidence interval for a population proportion. The purpose of this activity is to evaluate the performance of this alternative interval.

This activity presumes that you have already completed Activity 9.1. This activity uses the Shiny apps titled CIProportionVisualization and AlternativeConfidenceIntervalVisualization that are part of the collection of Shiny web apps that accompany this text. These apps can be found at **www.cengage.com/Peck**.

Background: In Section 9.2, it was suggested that a confidence interval of the form

$$\hat{p}_{\text{mod}} \pm (z \text{ critical value})\sqrt{\frac{\hat{p}_{\text{mod}}(1-\hat{p}_{\text{mod}})}{n}}$$

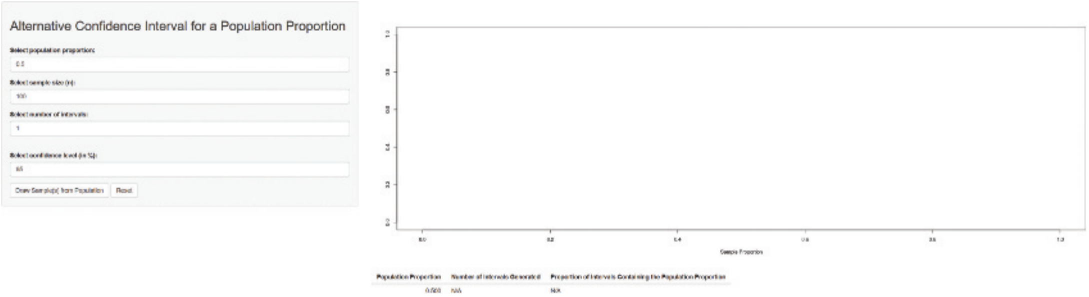
where $\hat{p}_{\text{mod}} = \dfrac{\text{successes} + 2}{n + 4}$ is an alternative to the usual large-sample z confidence interval. This alternative interval is preferred by many statisticians because, in repeated sampling, the proportion of intervals constructed that include the actual value of the population proportion, p, tends to be closer to the stated confidence level. In this activity, you will explore how the "capture rates" for the two different interval estimation methods compare.

Copyright 2025 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s). Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

Open the app CIProportionVisualization that was also used in Activity 9.1. Use this app to generate 1000 95% confidence intervals for samples of size $n = 40$ from a population with $p = 0.3$. Notice that $n = 40$ is large enough to satisfy the sample size conditions for the large-sample confidence interval for a population proportion ($np \geq 10$ and $n(1 - p) \geq 10$).

1. How does the proportion of intervals constructed that include $p = 0.3$, the value of the population proportion, compare to 0.95? Does this surprise you? Explain.

Now open the app Alternative Confidence Interval Visualization. This app is similar to the one used to generate large-sample confidence intervals, except this app selects a random sample and then uses the alternative method based on p_{mod} to generate the confidence intervals. You should see a screen similar to the one below.


Alternative Confidence Interval for a Population Proportion

Select population proportion:
Select sample size (n):
Select number of intervals:
Select confidence level (in %):
Draw Sample(s) from Population
Reset
Sample Proportion
Population Proportion
Number of Intervals Generated
Proportion of Intervals Containing the Population Proportion

2. Use the app to generate 1000 95% confidence intervals for samples of size $n = 40$ from a population with $p = 0.3$. How does the proportion of intervals constructed that include $p = 0.3$, the population proportion, compare to 0.95? Is this proportion closer to 0.95 than was the case for the large-sample z interval?
3. Experiment with different combinations of values of sample size and population proportion p. Can you find a combination for which the large sample z interval has a capture rate that is close to 95%? Can you find a combination for which it has a capture rate that is even farther from 95% than it was for $n = 40$ and $p = 0.3$? How does the modified interval perform in each of these cases?

Activity 9.3 Verifying Signatures on a Recall Petition

The purpose of this activity is to use a confidence interval to estimate the proportion of valid signatures on a recall petition.

Background

In 2020, petitions were submitted to the California Secretary of State calling for the recall of Governor Gavin Newsom. Each of California's 58 counties reported the number of valid signatures on the petitions from that county so that the state could determine whether there were enough to certify the recall and set a date for the recall election.

To verify that signatures on the recall petitions are valid, each county has the option of selecting a random sample of at least 5% of the signatures submitted for verification. The sample is then used to estimate the proportion of valid signatures submitted. Marin County reported receiving a total of 8640 signatures and checked the signatures in a random sample of size 2916. Of the 2916 signatures checked, 2651 were found to be valid (see elections.cdn.sos.ca.gov/recalls/cumulative-newsom-heatlie.pdf).

1. Use the data from the random sample of 2916 Marin County signatures to construct a 95% confidence interval for the proportion of Marin County petition signatures that are valid.
2. Given that there was a total of 8640 signatures from Marin County submitted, what do you think would be reasonable to report for the total number of valid signatures from Marin County?

Activity 9.4 A Meaningful Paragraph

The purpose of this activity is to allow you to demonstrate your understanding of key concepts introduced in this chapter.

Write a meaningful paragraph that includes the following six terms: **sample**, **population**, **confidence level**, **estimate**, **mean**, **margin of error**.

A "meaningful paragraph" is a coherent piece writing in an appropriate context that uses all of the listed words. The paragraph should show that you understand the meanings of the terms and their relationship to one another. A sequence of sentences that just define the terms is *not* a meaningful paragraph. When choosing a context, think carefully about the terms you need to use. Choosing a good context will make writing a meaningful paragraph easier.

Summary Key Concepts and Formulas

Term or Formula	Comment
Point estimate	A single number, based on sample data, that represents a plausible value of a population characteristic.
Unbiased statistic	A statistic that has a sampling distribution with a mean equal to the value of the population characteristic to be estimated.
Confidence interval	An interval calculated from sample data that provides a range of plausible values for a population characteristic.
Confidence level	A number that provides information on how much "confidence" you can have in the method used to construct a confidence interval estimate. The confidence level specifies the percentage of all possible samples that will produce an interval containing the value of the population characteristic.
$\hat{p} \pm (z \textbf{ critical value})\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$	A formula used to construct a confidence interval for p when the sample size is large.
$n = p(1-p)\left(\frac{1.96}{M}\right)^2$	A formula used to calculate the sample size necessary for estimating p to within an amount M with 95% confidence. (For other confidence levels, replace 1.96 with an appropriate z critical value.)
$\bar{x} \pm (z \textbf{ critical value})\frac{\sigma}{\sqrt{n}}$	A formula used to construct a confidence interval for μ when σ is known and either the sample size is large or the population distribution is normal.
$\bar{x} \pm (t \textbf{ critical value})\frac{s}{\sqrt{n}}$	A formula used to construct a confidence interval for μ when σ is unknown and either the sample size is large or the population distribution is normal.
$n = \left(\frac{1.96\sigma}{M}\right)^2$	A formula used to calculate the sample size necessary for estimating μ to within an amount M with 95% confidence. (For other confidence levels, replace 1.96 with an appropriate z critical value.)

Chapter Review Exercises 9.72–9.82

9.72 Suppose that county planners are interested in learning about the proportion of county residents who would pay a fee for a curbside recycling service if the county were to offer this service. Two different people independently selected random samples of county residents and used their sample data to construct the following confidence intervals for the proportion who would pay for curbside recycling:

Interval 1: (0.68, 0.74)
Interval 2: (0.68, 0.72)

a. Explain how it is possible that the two confidence intervals are not centered in the same place.

b. Which of the two intervals conveys more precise information about the value of the population proportion?

c. If both confidence intervals are associated with a 95% confidence level, which confidence interval was based on the smaller sample size? How can you tell?

d. If both confidence intervals were based on the same sample size, which interval has the higher confidence level? How can you tell?

9.73 Based on data from a representative sample of adult Americans (***USA Today*, August 4, 2021**), a 90% confidence interval for the proportion of adult Americans who said they believe that sea rise is a direct result of climate change is (0.503, 0.556). What is the meaning of the 90% confidence level associated with this interval?

9.74 Suppose that college students with a checking account typically write relatively few checks in any given month, whereas people who are not college students typically write many more checks during a month. Suppose that 50% of a bank's accounts are held by students and that 50% are held by people who are not college students. Let x represent the number of checks written in a given month by a randomly selected bank customer.

a. Give a sketch of what the population distribution of x might look like.

b. Suppose that the mean value of x is 22.0 and that the standard deviation is 16.5. If a random sample of $n = 100$ customers is to be selected and $\bar{x}$ denotes the sample mean number of checks written during a particular month, where is the sampling distribution of $\bar{x}$ centered, and what is the standard deviation of the sampling distribution of $\bar{x}$? Sketch a rough picture of the sampling distribution.

c. What is the approximate probability that $\bar{x}$ is at most 20? At least 25?

9.75 The report **"The Politics of Climate" (Pew Research Center, October 4, 2016, pewinternet.org)** summarized data from a survey on public opinion of renewable and other energy sources. It was reported that 52% of the people in a sample from Western states said that they have considered installing solar panels on their homes. This percentage was based on a representative sample of 369 homeowners in the Western United States. Use the given information to construct and interpret a 90% confidence interval for the proportion of all homeowners in Western states who have considered installing solar panels.

9.76 The report referenced in the previous exercise also indicated that 33% of those in a representative sample of 533 homeowners in Southern states said that they had considered installing solar panels.

a. Use the given information to construct and interpret a 90% confidence interval for the proportion of all homeowners in the Southern states who have considered installing solar panels.

b. Give two reasons why the confidence interval in Part (a) is narrower than the confidence interval calculated in the previous exercise.

9.77 Data from a survey of a representative sample was used to estimate that 32% of computer users in 2011 had tried to get on a Wi-Fi network that was not their own in order to save money **(*USA TODAY*, May 16, 2011)**. Because this estimate is from more than 10 years ago, you decide to conduct a survey to estimate this proportion for the current year. What is the required sample size if you want to estimate this proportion with a margin of error of 0.05? Calculate the required sample size first using 0.32 as a preliminary estimate of p and then using the conservative value of 0.5. How do the two sample sizes compare? What sample size would you recommend for this study?

9.78 A manufacturer of small appliances purchases plastic handles for coffeepots from an outside vendor. If a handle is cracked, it is considered defective and must be discarded. A large shipment of plastic handles is received. The proportion of defective handles p is of interest. How many handles from the shipment should be inspected to estimate p to within 0.1 with 95% confidence?

9.79 A manufacturer of college textbooks is interested in estimating the strength of the bindings produced by a particular binding machine. Strength can be measured by recording the force required to pull the pages from the binding. If this force is measured in pounds, how many books should be tested to estimate the mean force required to break the binding to within 0.1 pounds with 95% confidence? Assume that σ is known to be 0.8 pound.

9.80 The confidence intervals presented in this chapter give both lower and upper bounds on plausible values for the population characteristic being estimated. In some instances, only an upper bound or only a lower bound is appropriate. Using the same reasoning that gave the large sample interval in Section 9.3, you can say that when n is large, 99% of all samples have

$$\mu < \bar{x} + 2.33\frac{s}{\sqrt{n}}$$

(because the area under the z curve to the left of 2.33 is 0.99). Thus, $\bar{x} + 2.33\frac{s}{\sqrt{n}}$ is a 99% upper confidence bound for μ. For a sample of 539 patients waiting for bypass surgery, the mean wait time was 19 days and the standard deviation was 10 days. Calculate the 99% upper confidence bound for the population mean wait time for bypass patients.

9.81 When n is large, the statistic s is approximately unbiased for estimating σ and has approximately a normal distribution. The standard deviation of this statistic when the population distribution is normal is $\sigma_s \approx \frac{\sigma}{\sqrt{2n}}$ which can be estimated by $\frac{s}{\sqrt{2n}}$. An approximate large-sample confidence interval for the population standard deviation σ is then

$$s \pm (z \text{ critical value})\frac{s}{\sqrt{2n}}$$

For a random sample of 847 patients waiting for angiography, the standard deviation of wait time was 9 days. Calculate a 95% confidence interval for the population standard deviation of waiting time for angiography.

9.82 The interval from −2.33 to 1.75 captures an area of 0.95 under the z curve. This implies that another large-sample 95% confidence interval for μ has lower limit $\bar{x} - 2.33\frac{\sigma}{\sqrt{n}}$ and upper limit $\bar{x} + 1.75\frac{\sigma}{\sqrt{n}}$. Would you recommend using this 95% interval over the 95% interval $\bar{x} \pm 1.96\frac{\sigma}{\sqrt{n}}$ discussed in the text? Explain. (Hint: Look at the width of each interval.)

Copyright 2025 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s). Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

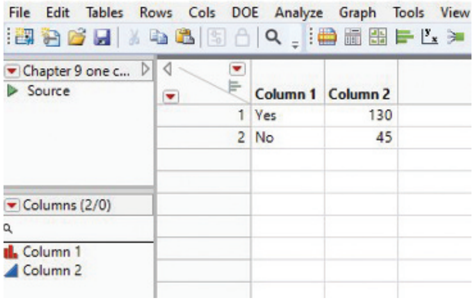
Technology Notes

Confidence Intervals for Proportions

JMP

Summarized data

1. Enter the data table into the JMP data table with categories in the first column and counts in the second column



	Column 1	Column 2
1	Yes	130
2	No	45

Source: JMP

2. Click **Analyze** and select **Distribution**
3. Click and drag the first column name from the box under **Select Columns** to the box next to **Y, Columns**
4. Click and drag the second column name from the box under **Select Columns** to the box next to **Freq**
5. Click **OK**
6. Click the red arrow next to the column name and click **Confidence Interval** then select the appropriate level or select **Other** to input a level that is not listed

Raw data

1. Enter the raw data into a column
2. Click **Analyze** and select **Distribution**
3. Click and drag the first column name from the box under **Select Columns** to the box next to **Y, Columns**
4. Click **OK**
5. Click the red arrow next to the column name and click **Confidence Interval** then select the appropriate level or select **Other** to input a level that is not listed

Minitab

Summarized data

1. Click **Stat** then click **Basic Statistics** then click **1 Proportion…**
2. Select **Summarized data** from the dropdown menu
3. In the box next to **Number of Trials:** type the value for *n*, the total number of trials
4. In the box next to **Number of events:** type the value for the number of successes
5. Click **Options…**
6. Input the appropriate confidence level in the box next to **Confidence Level**
7. Select **Normal approximation** from the dropdown menu next to **Method:**
8. Click **OK**
9. Click **OK**

Raw data

1. Input the raw data into a column
2. Click **Stat** then click **Basic Statistics** then click **1 Proportion…**
3. Select **One or more samples, each in a column** from the dropdown menu
4. Click on the box below the dropdown menu then double click the column name where the raw data is stored to copy the column name into the box
5. Click **Options…**
6. Input the appropriate confidence level in the box next to **Confidence Level**
7. Select **Normal approximation** from the dropdown menu next to **Method:**
8. Click **OK**
9. Click **OK**

SPSS

SPSS does not have the functionality to automatically produce confidence intervals for a population proportion.

Excel

Excel does not have the functionality to automatically produce confidence intervals for a population proportion. However, you can manually type in the formula for the lower and upper limits separately into two different cells to have Excel calculate the result for you.

TI-83/84

1. Press the **STAT** key
2. Highlight **TESTS**
3. Highlight **1-PropZInt** and press **ENTER**
4. Next to ***x*** type the number of successes
5. Next to ***n*** type the number of trials, *n*
6. Next to **C-Level** type the value for the confidence level
7. Highlight **Calculate** and press **ENTER**

TI-Nspire

1. Under **Scratchpad** select **A Calculate**
2. Press the **menu** key then select **6:Statistics** then select **6:Confidence Intervals** then **5:1-Prop *z* Interval…** then press **enter**
3. In the box next to **Successes, *x*** type the number of successes
4. In the box next to ***n*** type the number of trials, *n*
5. In the box next to **C Level** type the confidence level
6. Press **OK**

Confidence Interval for μ Based on *t*-distribution

JMP

1. Input the data into a column
2. Click **Analyze** and select **Distribution**
3. Click and drag the column name from the box under **Select Columns** to the box next to **Y, Response**
4. Click **OK**
5. Click the red arrow next to the column name and select **Confidence Interval** then select the appropriate confidence level or click **Other** to specify a level

Minitab

Summarized data

1. Click **Stat** then click **Basic Statistics** then click **1-Sample t…**

Copyright 2025 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s). Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

2. Select **Summarized data** from the dropdown menu
3. In the box next to **Sample size:** type the value for *n*, the sample size
4. In the box next to **Mean:** type the value for the sample mean
5. In the box next to **Standard deviation:** type the value for the sample standard deviation
6. Click **Options…**
7. Input the appropriate confidence level in the box next to **Confidence Level**
8. Click **OK**
9. Click **OK**

Raw data

1. Input the raw data into a column
2. Click **Stat** then click **Basic Statistics** then click **1-Sample t…**
3. Select **One or more samples, each in a column** from the dropdown menu
4. Click on the box below the dropdown menu then double click the column name where the raw data is stored to copy the column name into the box
5. Click **Options…**
6. Input the appropriate confidence level in the box next to **Confidence Level**
7. Click **OK**
8. Click **OK**

SPSS

1. Input the data into a column
2. Click **Analyze** then select **Compare Means** then select **One-Sample T Test…**
3. Highlight the column name for the variable
4. Click the arrow to move the variable to the **Test Variable(s):** box
5. Click **Options…**
6. Input the appropriate confidence level in the box next to **Confidence Interval Percentage:**
7. Click **Continue**
8. Click **OK**

Note: The confidence interval for the population means is in the **One-Sample Test** table.

Excel

Note: Excel does not have the functionality to produce a confidence interval for a single population mean automatically. However, you can use Excel to produce the estimate for the sample mean and the margin of error for the confidence interval using the steps below.

1. Input the raw data into a column
2. Click the **Data** ribbon
3. Click **Data Analysis** in the **Analysis** group

   Note: If you do not see **Data Analysis** listed on the Ribbon, see the Technology Notes for Chapter 2 for instructions on installing this add-on.

4. Select **Descriptive Statistics** from the dialog box and click **OK**
5. Click in the box next to **Input Range:** and select the data (if you used a column title, check the box next to **Labels in first row**)
6. Click the box next to **Confidence Level for Mean:**
7. In the box to the right of **Confidence Level for Mean:** type in the confidence level you are using
8. Click **OK**

Note: The margin of error can be found in the row titled Confidence Level. In order to find the lower limit of the confidence interval, subtract this from the mean shown in the output. To find the upper limit of the confidence interval, add this to the mean shown in the output.

TI-83/84

Summarized data

1. Press **STAT**
2. Highlight **TESTS**
3. Highlight **TInterval…** and press **ENTER**
4. Highlight **Stats** and press **ENTER**
5. Next to $\bar{\mathbf{x}}$ input the value for the sample mean
6. Next to **Sx** input the value for the sample standard deviation
7. Next to **n** input the value for the sample size
8. Next to **C-Level** input the appropriate confidence level
9. Highlight **Calculate** and press **ENTER**

Raw data

1. Enter the data into **L1** (In order to access lists press the **STAT** key, highlight the option called **Edit…** then press **ENTER**)
2. Press **STAT**
3. Highlight **TESTS**
4. Highlight **TInterval…** and press **ENTER**
5. Highlight **Data** and press **ENTER**
6. Next to **C-Level** input the appropriate confidence level
7. Highlight **Calculate** and press **ENTER**

TI-Nspire

Summarized data

1. Under **Scratchpad** select **A Calculate**
2. Press the **menu** key and select **6:Statistics** then **6:Confidence Intervals** then **2:t Interval…** and press **enter**
3. From the drop-down menu select **Stats**
4. Press **OK**
5. Next to $\bar{\mathbf{x}}$ input the value for the sample mean
6. Next to **Sx** input the value for the sample standard deviation
7. Next to **n** input the value for the sample size
8. Next to **C Level** input the appropriate confidence level
9. Press **OK**

Raw data

1. Enter the data into a data list (In order to access data lists select the spreadsheet option and press **enter**)

Note: Be sure to title the list by selecting the top row of the column and typing a title.

2. Press the **menu** key and select **4:Statistics** then **3:Confidence Intervals** then **2:t Interval…** and press **enter**
3. From the drop-down menu select **Data**
4. Press **OK**
5. Next to **List** select the list containing your data
6. Next to **C-Level** input the appropriate confidence level
7. Press **OK**

Chapter 10

Hypothesis Testing Using a Single Sample

Fizkes/Shutterstock.com

Chapter 9 considered situations in which the primary goal was to estimate the unknown value of some population characteristic. Sample data can also be used to decide whether a claim or *hypothesis* about a population characteristic is plausible.

For example, how do employers rate the career readiness of college students? Is there evidence that less than half of employers would rate college seniors as proficient in oral and written communication? The article **"Overconfident Students, Dubious Employers" (insidehighered.com, February 23, 2018)** summarizes results from a survey of a representative sample of 201 employers and 4213 graduating seniors. Of the 201 employers surveyed, only 84 rated college seniors as proficient in oral and written communication. With p representing the proportion of all employers who would rate college seniors as proficient, the hypothesis testing methods of this chapter can be used to decide whether the sample data provide convincing evidence that p is less than 0.50.

As another example, a report released by the National Association of Colleges and Employers stated that the average starting salary for students graduating with a bachelor's degree in 2021 was \$58,862 **("What Students Want in a Job: Security, Skill Development, Friendly Coworkers, and Good Benefits," October 11, 2022, naceweb.org)**. Suppose that you are interested in investigating whether the mean starting salary for students graduating from your university this year is greater than the 2021 average of \$58,862. You select a random sample of $n = 40$ graduates from the graduating class of your university and determine the starting salary of each one.

If this sample produced a mean starting salary of \$59,136 and a standard deviation of \$1214, is it reasonable to conclude that μ, the mean starting salary for all graduates in the graduating class at your university, is greater than 58,862? In this chapter, you will see how the sample data can be used to decide whether $\mu > 58{,}862$ is a reasonable conclusion.

Learning Objectives

After completing this chapter, you will be able to:

- Translate a research question into a null and alternative hypothesis.
- Describe Type I and Type II errors in context.
- Select an appropriate significance level based on consideration of the consequences of Type I and Type II errors.

Copyright 2025 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s). Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

- Explain why rejecting the null hypothesis implies strong support for the alternative hypothesis but failing to reject the null hypothesis does not imply strong support for the null hypothesis.
- Carry out a large-sample z test for a population proportion and interpret the results in context.
- Carry out a t test for a population mean and interpret the results in context.
- Describe the effect of the significance level and the sample size on the power of a test.
- Carry out a randomization test for a population proportion. (Optional)
- Carry out an exact binomial test for a population proportion. (Optional)
- Carry out a randomization test for a population mean. (Optional)

Section 10.1 Hypotheses and Test Procedures

A hypothesis is a claim or statement about the value of a single population characteristic or the values of several population characteristics. The following are examples of legitimate hypotheses:

$\mu = 1000$, where μ is the mean number of characters in an e-mail message
$p < 0.01$, where p is the proportion of e-mail messages that are undeliverable

In contrast, the statements $\bar{x} = 1000$ and $\hat{p} = 0.01$ are *not* hypotheses, because $\bar{x}$ and $\hat{p}$ are *sample* characteristics.

A test of hypotheses is a method that uses sample data to decide between two competing claims (hypotheses) about a population characteristic. One hypothesis might be $\mu = 1000$ and the other $\mu \neq 1000$ or one hypothesis might be $p = 0.01$ and the other $p < 0.01$. If it were possible to carry out a census of the entire population, you would know which of the two hypotheses is correct, but usually you must decide between them using information from a sample.

A criminal trial is a familiar situation in which a choice between two contradictory claims must be made. The person accused of a crime must be judged either guilty or not guilty. Under the U.S. system of justice, the individual on trial is initially presumed not guilty. Only strong evidence to the contrary causes the not guilty claim to be rejected in favor of a guilty verdict. The burden is on the prosecution to provide convincing evidence that supports the guilty claim. In some other countries, the perspective in criminal proceedings is the opposite. Once enough evidence has been presented to justify bringing an individual to trial, the initial assumption is that the accused is guilty. The burden then falls on the accused to provide convincing evidence to support a not guilty claim.

As in a judicial proceeding, you will initially assume that a particular hypothesis, called the **null hypothesis**, is the correct one. You then consider the evidence (the sample data) and reject the null hypothesis in favor of the competing hypothesis, called the **alternative hypothesis**, only if there is *convincing* evidence against the null hypothesis.

Definitions

Null hypothesis: A claim about a population characteristic that is initially assumed to be true. The null hypothesis is denoted by H_0.

Alternative hypothesis: A competing claim about a population characteristic. The alternative hypothesis is denoted by H_a.

In carrying out a test of H_0 versus H_a, the null hypothesis H_0 will be rejected in favor of H_a only if the sample provides convincing evidence that H_0 is false.

If the sample does not provide such evidence, H_0 will not be rejected.

The two possible conclusions in a test of hypotheses are *reject* H_0 or *fail to reject* H_0.

Example 10.1 Tennis Ball Diameters

Understand the context ➤

The International Tennis Federation requires that tennis balls have a diameter between 2.57 and 2.70 inches. Because of variation in the manufacturing process, tennis balls produced by a particular machine do not all have identical diameters. You can use μ to represent the mean diameter for all tennis balls produced by this machine. Suppose that the machine was initially calibrated to achieve the design specification $\mu = 2.64$ inches. However, the manufacturer is now concerned that the diameters no longer conform to this specification. If sample evidence suggests that $\mu \neq 2.64$ inches, the production process will be halted while the machine is recalibrated. Because stopping production is costly, the manufacturer wants to be quite sure that $\mu \neq 2.64$ inches before undertaking recalibration. Under these circumstances, a sensible choice of hypotheses is

H_0: $\mu = 2.64$ (the specification is being met, so recalibration is unnecessary)
H_a: $\mu \neq 2.64$ (the specification is not being met, so recalibration is necessary)

H_0 would be rejected in favor of H_a only if the sample provides convincing evidence against the null hypothesis.

Example 10.2 Compact Florescent Lightbulb Lifetimes

Understand the context ➤

Compact fluorescent (cfl) lightbulbs are more energy efficient than standard incandescent light bulbs. Ecobulb brand 60-watt cfl lightbulbs state on the package “Average life 8,000 hours.” Using μ to denote the true mean life of Ecobulb 60-watt cfl light bulbs, the advertised claim is $\mu = 8000$ hours. People who purchase this brand would be unhappy if μ is actually less than the advertised value.

Suppose that a sample of Ecobulb cfl lightbulbs is selected and tested. The lifetime for each bulb in the sample is recorded. The sample data can then be used to test the hypothesis $\mu = 8000$ hours against the hypothesis $\mu < 8000$ hours. The accusation that the company is overstating the mean lifetime is a serious one, and it is reasonable to require convincing evidence before concluding that $\mu < 8000$. This suggests that the claim $\mu = 8000$ should be selected as the null hypothesis and that $\mu < 8000$ should be selected as the alternative hypothesis. Then

H_0: $\mu = 8000$

would be rejected in favor of

H_a: $\mu < 8000$

only if sample provides convincing evidence that the initial assumption, $\mu = 8000$ hours, is not plausible.

Because the alternative hypothesis in Example 10.2 is $\mu < 8000$ (the actual mean lifetime is less than the advertised value), it might have seemed sensible to state H_0 as the inequality $\mu \geq 8000$. The claim $\mu \geq 8000$ is in fact the *implicit* null hypothesis, but it is common to state H_0 as a claim of equality. There are several reasons for this. First, the development of a decision rule is most easily understood if there is only a single hypothesized value of μ (or p or whatever other population characteristic is under consideration). Second, suppose that the sample data provided compelling evidence that H_0: $\mu = 8000$ should be rejected in favor of H_a: $\mu < 8000$. This means that you were convinced by the sample data that the population mean was smaller than 8000. It follows that you would have also been convinced that the population mean could not have been 8001 or 8010 or any other value that was greater than 8000. This means that the conclusion when testing H_0: $\mu = 8000$ versus H_a: $\mu < 8000$ is always the same as the conclusion for a test where the null hypothesis is H_0: $\mu \geq 8000$. For these reasons, the null hypothesis H_0 is usually stated as a claim of equality.

The form of a null hypothesis is

$$H_0\text{: population characteristic} = \text{hypothesized value}$$

where the hypothesized value is a specific number determined by the problem context.

The alternative hypothesis will have one of the following three forms:

H_a: population characteristic $>$ hypothesized value

H_a: population characteristic $<$ hypothesized value

H_a: population characteristic $\neq$ hypothesized value

What Do You Think?

Based on the information in the preceding box, would you be able to test a null hypothesis of H_0: $\mu = 17$ versus an alternate hypothesis of H_a: $\mu \neq 20$?

Notice that you can test H_0: $p = 0.1$ versus H_a: $p < 0.1$, but you can't test H_0: $\mu = 50$ versus H_a: $\mu > 100$. The number appearing in the alternative hypothesis must be the same as the hypothesized value in H_0.

Example 10.3 illustrates how the selection of H_0 (the claim initially assumed to be true) and H_a depends on the objectives of a study.

Example 10.3 Evaluating a New Medical Treatment

Understand the context ➤

A medical research team has been given the task of evaluating a new laser treatment for certain types of tumors. Consider the following two scenarios:

Scenario 1: The current standard treatment is considered reasonable and safe by the medical community, has no major side effects, and has a known success rate of 0.85 (85%).

Scenario 2: The current standard treatment sometimes has serious side effects, is costly, and has a known success rate of 0.30 (30%).

In the first scenario, research efforts would probably be directed toward determining whether the new treatment has a greater success rate than the standard treatment. Unless there is convincing evidence of this, it is unlikely that current medical practice would change. With p representing the true proportion of successes for the new laser treatment, the following hypotheses would be tested:

$$H_0\text{: } p = 0.85 \quad \text{versus} \quad H_a\text{: } p > 0.85$$

In this case, rejecting the null hypothesis indicates convincing evidence that the success rate is greater for the new treatment.

In the second scenario, the current standard treatment does not have much to recommend it. The new laser treatment may be considered preferable because of cost or because it has fewer or less serious side effects, as long as the success rate for the new procedure is no worse than the success rate of the standard treatment. Here, researchers might decide to test the hypotheses

$$H_0\text{: } p = 0.30 \quad \text{versus} \quad H_a\text{: } p < 0.30$$

If the null hypothesis is rejected, the new treatment will not be recommended as an alternative to the standard treatment, because there is convincing evidence that the laser method has a lower success rate.

If the null hypothesis is not rejected, you are able to conclude only that there is not convincing evidence that the success rate for the laser treatment is less than the success rate

Copyright 2025 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s). Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

for the standard treatment. This is *not* the same as saying that you have evidence that the laser treatment is as good as the standard treatment. If medical practice were to embrace the new procedure, it would not be because it has a greater success rate but rather because it costs less or has fewer side effects, and there is not convincing evidence that it has a lower success rate than the standard treatment.

A statistical hypothesis test is only capable of demonstrating strong support for the alternative hypothesis (by rejecting the null hypothesis). When the null hypothesis is not rejected, it does not mean strong support for the null hypothesis—only that there is not *convincing* evidence against it.

In the lightbulb scenario of Example 10.2, if H_0: $\mu = 8000$ is rejected in favor of H_a: $\mu < 8000$, it would be because you have convincing evidence that actual mean lifetime is less than the advertised value. However, not rejecting H_0 does not necessarily provide strong support for the advertised claim. If the objective is to demonstrate that the mean lifetime is greater than 8000 hours, the hypotheses that would be tested are H_0: $\mu = 8000$ versus H_a: $\mu > 8000$. Then rejecting H_0 indicates convincing evidence that $\mu > 8000$. When deciding which alternative hypothesis to use, keep the research objectives in mind.

Exercises 10.1–10.11

10.1 Explain why the statement $\bar{x} = 50$ is not a legitimate hypothesis.

10.2 For the following pairs, indicate which do not comply with the rules for setting up hypotheses, and explain why:

a. H_0: $\mu = 15$, H_a: $\mu = 15$
b. H_0: $p = 0.4$, H_a: $p > 0.6$
c. H_0: $\mu = 123$, H_a: $\mu < 123$
d. H_0: $\mu = 123$, H_a: $\mu = 125$
e. H_0: $\hat{p} = 0.1$, H_a: $\hat{p} \neq 0.1$

10.3 To determine whether the pipe welds in a nuclear power plant meet specifications, a random sample of welds is selected and tests are conducted on each weld in the sample. Weld strength is measured as the force required to break the weld. Suppose that the specifications state that the mean strength of welds should exceed 100 lb/in^2. The inspection team decides to test H_0: $\mu = 100$ versus H_a: $\mu > 100$. Explain why this alternative hypothesis was chosen rather than $\mu < 100$.

10.4 According to an article in *Science Daily* **("Still No Strong Evidence That Adjunct Treatment with HGH in IVF Improves Results," sciencedaily.com, July 4, 2016)**, human growth hormone (HGH) is sometimes used in combination with in vitro fertilization (IVF) to increase the likelihood of a successful pregnancy. A large study found that "there was no strong evidence" that the proportion of successful pregnancies while taking HGH along with IVF was greater than the success rate for IVF alone.

a. Is this consistent with testing

H_0: HGH in addition to IVF increases the chance of getting pregnant

versus

H_a: HGH in addition to IVF does not increase the chance of getting pregnant

or with testing

H_o: HGH in addition to IVF does not increase the chance of getting pregnant

versus

H_a: HGH in addition to IVF increases the chance of getting pregnant

Explain.

b. Does the stated conclusion of "no strong evidence" indicate that the null hypothesis was rejected? Explain.

10.5 A press release about a paper that appeared in ***The Journal of Youth and Adolescence*** **(springer.com, August 26, 2013)** was titled "Video Games Do Not Make Vulnerable Teens More Violent." The press release includes the following statement about the study described in the paper: "Study finds no evidence that violent video games increase antisocial behavior in youths with preexisting psychological conditions." In the context of a hypothesis test with the null hypothesis being that video games do not increase antisocial

Copyright 2025 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s). Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

behavior, explain why the title of the press release is misleading.

10.6 Many companies conduct background checks for job candidates, but some companies also continue to conduct background checks even after a person has been hired and is working for the company. A 2021 survey conducted by the Professional Background Screening Association found that 19% of the companies surveyed reported that they carried out post-hire background checks **("Employment Background Checks Shift to Continuous," techtarget.com, March 21, 2022)**. Suppose that you are interested in using the data from this survey to determine if there is convincing evidence that more than 15% of companies carry out post-hire background checks. What pair of hypotheses would you test?

10.7 A national survey of 1023 adult Americans conducted by Gallup **("Americans' Views of Their Financial Situation Worsen," February 3, 2021)** asked survey participants if they thought they were in better financial shape than they were 1 year ago. Suppose that you want to determine if the survey data provide convincing evidence that a majority of adult Americans believe they are in better financial shape than 1 year ago. With $p =$ the population proportion of adult Americans who believe they are better off, what hypotheses should you test?

10.8 A researcher speculates that because of differences in diet, Japanese children may have a lower mean blood cholesterol level than U.S. children. Suppose that the mean level for U.S. children is known to be 170. Let μ represent the mean blood cholesterol level for all Japanese children. What hypotheses should the researcher test?

10.9 A county commissioner must vote on a resolution that would commit substantial resources to the construction of a sewer in an outlying residential area. Their fiscal decisions have been criticized in the past, so the commissioner decides to take a survey of people in their district to find out whether they favor spending money for a sewer system. The commissioner will vote to appropriate funds only if they are convinced that a majority (more than 50%) of the people in the district favor the measure. What hypotheses should the commissioner test?

10.10 A cruise ship charges passengers $5 for a can of soda. Because of passenger complaints, the ship manager has decided to try out a plan with a lower price. The manager thinks that with a lower price, more cans will be sold, which would mean that the ship would still make a reasonable total profit. With the old pricing, the mean number of cans sold per passenger for a 10-day trip was 10.3 cans.

Suppose μ represents the mean number of cans per passenger for the new pricing. What hypotheses should the ship manager test if they want to determine if the mean number of cans sold is greater for the new pricing plan?

10.11 The article **"Facebook Use and Academic Performance Among College Students,"** ***Computers in Human Behavior*** **[2015]: 265–272)** estimated that 70 percent of students at a large public university in California who are Facebook users log into their Facebook profiles at least six times a day. Suppose that you plan to select a random sample of 400 students at your college. You will ask each student in the sample if they are a Facebook user and if they log into their Facebook profile at least six times a day. You plan to use the resulting data to decide if there is evidence that the proportion for your college is different from the dated proportion reported in the article for the college in California. What hypotheses should you test?

Section 10.2 Errors in Hypothesis Testing

Once hypotheses have been formulated, a **test procedure** uses sample data to determine whether the null hypothesis, H_0, should be rejected. Just as a jury may reach the wrong verdict in a trial, there is some chance that using a test procedure with sample data may lead you to the wrong conclusion about a population characteristic. This section introduces the types of errors that can occur and looks at how the choice of a test procedure influences the chances of these errors.

One incorrect conclusion in a criminal trial is for a jury to convict an innocent person. Another is for a guilty person to be set free. Similarly, there are two different types of errors that might be made when making a decision in a hypothesis test. One type of error involves rejecting H_0 even though the null hypothesis is true. The second type of error results from failing to reject H_0 when it is false. These errors are known as Type I and Type II errors, respectively.

Copyright 2025 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s). Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

Definitions

Type I error: The error of rejecting H_0 when H_0 is true

Type II error: The error of failing to reject H_0 when H_0 is false

The only way to guarantee that neither type of error occurs is to base the decision on a census of the entire population. Risk of error is the price paid for basing the decision on sample data.

Example 10.4 On-Time Arrivals

Understand the context ➤

The **U.S. Bureau of Transportation Statistics** reports that for March of 2023, 75.4% of all domestic passenger flights arrived within 15 minutes of the scheduled arrival time (***Air Travel Consumer Reports,*** **May, 2023)**. Flights that arrive within 15 minutes of the scheduled time are considered to be on-time. Suppose that an airline with a poor on-time record decides to offer its employees a bonus if the airline's proportion of on-time flights exceeds the overall industry rate of 0.754 in an upcoming month. You can use p to represent the actual proportion of the airline's flights that are on time during the month of interest. A random sample of flights might be selected and used as a basis for choosing between

$$H_0\text{: } p = 0.754 \quad \text{and} \quad H_a\text{: } p > 0.754$$

In this context, a Type I error (rejecting a true H_0) is concluding that the airline on-time rate exceeds the overall industry rate, when in fact the airline does not have a better record. This Type I error would result in the airline rewarding its employees when the actual proportion of on-time flights was not actually greater than 0.754. A Type II error (not rejecting a false H_0) is not concluding that the airline's on-time proportion is greater than the industry proportion when the airline really did have a better on-time record. A type II error would result in the airline employees *not* receiving a reward that they deserved. Notice that the consequences associated with Type I and Type II errors are different.

Example 10.5 Yoga and Exercise to Reduce Blood Pressure

Understand the context ➤

The authors of the paper **"Impact of Yoga on Global Cardiovascular Risk as an Add-On to a Regular Exercise Regimen in Patients with Hypertension" (*Canadian Journal of Cardiology* [2023]: 57–62)** hypothesized that adding 30 minutes of yoga to each exercise session would lower blood pressure of people diagnosed with hypertension (high blood pressure) who exercised regularly. If this is the case, it might represent an alternative to trying to control blood pressure though medication. The researchers carried out an experiment designed to test the claim that yoga, along with regular exercise, could lower blood pressure more than just exercise alone. Using μ to represent the mean decrease in blood pressure for people with hypertension who exercise regularly and who add 30 minutes of yoga to their exercise sessions, data from the experiment could be used to test

$$H_0\text{: } \mu = 0$$

versus

$$H_a\text{: } \mu > 0$$

The null hypothesis states that adding yoga to regular exercise is not effective in reducing blood pressure—that the mean decrease in blood pressure is 0. The alternative hypothesis states that adding yoga to regular exercise is effective in reducing blood pressure—that the decrease is greater than 0.

In this context, a Type I error would be to incorrectly conclude that adding yoga is effective in reducing blood pressure. A potential consequence of making a Type I error is that doctors would start recommending the addition of yoga, even though it really did not contribute to reducing blood pressure.

A Type II error would be not concluding that the addition of yoga was beneficial when adding yoga really did result in a reduction in blood pressure. A potential consequence of making a Type II error is that a potential nonmedical treatment for hypertension would be overlooked.

The accompanying box introduces the terminology and notation used to describe error probabilities.

Definitions

The **probability of a Type I error** is denoted by $\boldsymbol{\alpha}$ and is called the **significance level** of the test. For example, a test with $\boldsymbol{\alpha} = 0.01$ is said to have a significance level of 0.01.

The **probability of a Type II error** is denoted by $\boldsymbol{\beta}$.

Example 10.6 Early Detection of Lung Cancer

Understand the context ➤

Early detection has been shown to increase the chance of survival for patients with lung cancer. The paper **"Urinary Protein Biomarkers in the Early Detection of Lung Cancer" (*Cancer Prevention Research* [2015]: 111–117)** includes data from a study to determine if it is possible to accurately diagnose lung cancer using biomarkers found in a patient's urine. The researchers developed a screening method and then tested the method with a group of people who were known to have lung cancer and a group of patients who were known to not have lung cancer. The paper included the following information:

- The test correctly identified lung cancer in 39 of 54 patients known to have lung cancer.
- The test correctly identified as cancer-free all of the 49 people tested who were known not to have lung cancer.

You can think of using this test to choose between two hypotheses:

H_0: patient has lung cancer
H_a: patient does not have lung cancer

Although these are not "statistical hypotheses" (statements about a population characteristic), the possible decision errors are analogous to Type I and Type II errors.

In this situation, believing that a patient with lung cancer is cancer-free would be a Type I error—rejecting the hypothesis of lung cancer when it is actually true. Believing that a cancer-free patient does have lung cancer is a Type II error—not rejecting the null hypothesis when it is actually false.

Based on the study results, you can estimate the error probabilities. The probability of a Type I error, α, is approximately $(54 - 39)/54 = 0.278$. The probability of a Type II error, β, is approximately $0/49 = 0$.

An ideal test procedure would result in both $\alpha = 0$ and $\beta = 0$. However, if a decision must be based on incomplete information—a sample rather than a census—it is impossible to achieve this ideal. The standard test procedures will allow you to control α, but they provide no direct control over β. Because α represents the probability of rejecting a true null hypothesis, selecting a significance level $\alpha = 0.05$ results in a test procedure that, used over and over with different random samples, rejects a *true* H_0 about 5 times in 100. Selecting $\alpha = 0.01$ results in a test procedure with a Type I error rate of 1% in long-term repeated use. Choosing a small value for α implies that you want a procedure for which the risk of a Type I error is quite small.

One question arises naturally at this point: If you can select α, the probability of making a Type I error, why would you ever select $\alpha = 0.05$ rather than $\alpha = 0.01$? Why not always select a very small value for α? To achieve a small probability of making a Type I error, you

Copyright 2025 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s). Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

would need the corresponding test procedure to require the evidence against H_0 to be very strong before the null hypothesis can be rejected. Although this makes a Type I error unlikely, it increases the risk of a Type II error (*not* rejecting H_0 when it should have been rejected). The choice will depend on the consequences of Type I and Type II errors. If a Type II error has serious consequences, it may be a good idea to select a somewhat larger value for α.

In general, there is a compromise between small α and small β, leading to the following widely accepted principle for selecting the significance level to be used in a hypothesis test.

After assessing the consequences of Type I and Type II errors, identify the largest α that is acceptable. Then use a test procedure with this maximum acceptable value as the level of significance (because using a smaller α increases β).

Example 10.7 Lead in Tap Water

Understand the context ➤

The Environmental Protection Agency (EPA) has adopted what is known as the Lead and Copper Rule, which defines drinking water as unsafe if the concentration of lead is 15 parts per billion (ppb) or greater or if the concentration of copper is 1.3 parts per million (ppm) or greater.

With μ denoting the mean concentration of lead, the manager of a community water system might use lead level measurements from a sample of water specimens to test

$$H_0\text{: } \mu = 15 \quad \text{versus} \quad H_a\text{: } \mu < 15$$

The null hypothesis (which also implicitly includes the $\mu > 15$ case) states that the mean lead concentration is excessive by EPA standards. The alternative hypothesis states that the mean lead concentration is at an acceptable level and that the water system meets EPA standards for lead.

What Do You Think?

Given these hypotheses, what is a Type I error? What is a Type II error?

In this context, a Type I error leads to the conclusion that a water source meets EPA standards for lead when in fact it does not. Possible consequences of this type of error include health risks associated with excessive lead consumption (for example, increased blood pressure, hearing loss, and, in severe cases, anemia and kidney damage).

A Type II error is concluding that the water does not meet EPA standards for lead when it actually does. Possible consequences of a Type II error include elimination of a community water source. Because a Type I error might result in potentially serious public health risks, a small value of α (Type I error probability), such as $\alpha = 0.01$, could be selected. Of course, selecting a small value for α increases the risk of a Type II error. If the community has only one water source, a Type II error could also have very serious consequences for the community, and you might want to rethink the choice of α.

Exercises 10.12–10.22

10.12 Researchers at Boston's Children's Hospital and Harvard Medical School analyzed records of breast cancer screening and diagnostic evaluations **("National Expenditure for False-Positive Mammograms and Breast Cancer Overdiagnoses Estimated at $4 Billion a Year," *Health Affairs* [2015]: 576–583)**. Discussing the downsides of the screening process, the article states that the rate of false-positives is higher than previously thought, and that false-positives lead to unnecessary medical follow-up that can be costly.

Suppose that screening is used to decide between a null hypothesis of

H_0: no cancer is present

and an alternative hypothesis of

H_a: cancer is present.

Copyright 2025 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s). Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

(Although these are not hypotheses about a population characteristic, this exercise illustrates the definitions of Type I and Type II errors.) (Hint: See Example 10.6.)

a. Would a false-positive (thinking that cancer is present when in fact it is not) be a Type I error or a Type II error?

b. Describe a Type I error in the context of this problem, and discuss the possible consequences of making a Type I error.

c. Describe a Type II error in the context of this problem, and discuss the possible consequences of making a Type II error.

d. Which type of error are the researchers concerned about when they say that false-positives lead to unnecessary medical follow-up? Explain why it would be reasonable to use a small significance level.

10.13 The paper **"Breast MRI as an Adjunct to Mammography for Breast Cancer Screening in High-Risk Patients" (*American Journal of Roentgenology* [2015]: 889–897)** describes a study that investigated the usefulness of MRI (magnetic resonance imaging) to diagnose breast cancer. MRI exams of 650 people were reviewed. Of the 650 people, 13 had breast cancer, and the MRI exam detected breast cancer in 12 of these people. Of the 637 people who did not have breast cancer, the MRI correctly identified that no cancer was present for 547 of them. The accompanying table summarizes this information.

	Breast Cancer Present	**Breast Cancer Not Present**	**Total**
MRI Indicated Breast Cancer	12	90	102
MRI Did Not Indicate Breast Cancer	1	547	548
Total	13	637	650

Suppose that an MRI exam is used to decide between the two hypotheses

H_0: A person does not have breast cancer
H_a: A person has breast cancer

(Although these are not hypotheses about a population characteristic, this exercise illustrates the definitions of Type I and Type II errors.)

a. One possible error would be deciding that a person who has breast cancer is cancer-free. Is this a Type I error or a Type II error? Use the information in the table to approximate the probability of this type of error.

b. There is a second type of error that is possible in this context. Describe this error and use the information in the table to approximate the probability of this type of error.

10.14 Medical personnel are required to report suspected cases of child abuse. Because some diseases have symptoms that mimic those of child abuse, doctors who see a child with these symptoms must decide between two competing hypotheses:

H_0: symptoms are due to child abuse
H_a: symptoms are due to disease

(Although these are not hypotheses about a population characteristic, this exercise illustrates the definitions of Type I and Type II errors.) The article **"Blurred Line Between Illness, Abuse Creates Problem for Authorities" (*Macon Telegraph*, February 28, 2000)** included the following quote from a doctor in Atlanta regarding the consequences of making an incorrect decision: "If it's disease, the worst you have is an angry family. If it is abuse, the other kids (in the family) are in deadly danger."

a. For the given hypotheses, describe Type I and Type II errors.

b. Based on the quote regarding consequences of the two kinds of error, which type of error does the doctor quoted consider more serious? Explain.

10.15 How accurate are DNA tests used to identify animals from scat (animal droppings) samples? Researchers were interested in a DNA test to determine if non-native fox species were present in parts of New Zealand **("An Examination of the Accuracy of a Sequential PCR and Sequencing Test to Detect the Incursion of an Invasive Species: The Case of the Red Fox in Tasmania," *Journal of Applied Ecology* [2015]: 562–570)**. To evaluate the accuracy of the test, the test was used on 319 scat samples that were known to be from foxes and 208 scat samples that were known to be from other animals. The test correctly identified 298 of the 319 fox samples as being from foxes and incorrectly identified 9 of the 208 scat samples from other animals as being from foxes.

a. Consider using the result of this test to decide between the following two hypotheses:

H_0: scat is not from a fox
H_a: scat is from a fox

In the context of this problem, describe Type I and Type II errors. (Although these are not hypotheses about a population characteristic, this exercise illustrates the definitions of Type I and Type II errors.)

b. Based on the information given, what are the estimated values of α, the probability of a Type I error, and β, the probability of a Type II error?

10.16 A television manufacturer claims that (at least) 90% of its TV sets will not need service during the first 3 years of operation. A consumer agency wishes to check this claim, so it obtains a random sample of

Copyright 2025 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s). Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

$n = 100$ purchasers and asks each whether the set purchased needed repair during the first 3 years after purchase. Let $\hat{p}$ be the sample proportion of responses indicating no repair (so that no repair is identified with a success). Let p denote the actual proportion of successes for all sets made by this manufacturer.

The agency does not want to claim false advertising unless sample evidence strongly suggests that $p < 0.90$. The appropriate hypotheses are then H_0: $p = 0.90$ versus H_a: $p < 0.90$.

a. In the context of this problem, describe Type I and Type II errors, and discuss the possible consequences of each.

b. Would you recommend a test procedure that uses $\alpha = 0.10$ or one that uses $\alpha = 0.01$? Explain.

10.17 A manufacturer of hand-held calculators receives large shipments of printed circuits from a supplier. It is too costly and time-consuming to inspect all incoming circuits, so when each shipment arrives, a sample is selected for inspection. Information from the sample is then used to test H_0: $p = 0.01$ versus H_a: $p > 0.01$, where p is the actual proportion of defective circuits in the shipment.

If the null hypothesis is not rejected, the shipment is accepted, and the circuits are used in the production of calculators. If the null hypothesis is rejected, the entire shipment is returned to the supplier because of inferior quality. (A shipment is defined to be of inferior quality if it contains more than 1% defective circuits.)

a. In this context, define Type I and Type II errors.

b. From the calculator manufacturer's point of view, which type of error is considered more serious?

c. From the printed circuit supplier's point of view, which type of error is considered more serious?

10.18 Water specimens are taken from water used for cooling as it is being discharged from a power plant into a river. It has been determined that as long as the mean temperature of the discharged water is at most 150°F, there will be no negative effects on the river's ecosystem. To investigate whether the plant is in compliance with regulations that prohibit a mean discharge water temperature above 150°F, temperature will be determined for 50 water specimens at randomly selected times. The resulting data will be used to test the hypotheses H_0: $\mu = 150°\text{F}$ versus H_a: $\mu > 150°\text{F}$.

a. In the context of this example, describe Type I and Type II errors.

b. Which type of error would you consider more serious? Explain.

10.19 Suppose that for a particular hypothesis test, the consequences of a Type I error are very serious. Would you want to carry out the test using a small significance level α (such as 0.01) or a larger significance level (such as 0.10)? Explain the reason for your choice. (Hint: See discussion just before Example 10.7.)

10.20 Suppose that you are an inspector for the Fish and Game Department and that you are given the task of determining whether to prohibit fishing along part of the Oregon coast. You will close an area to fishing if it is determined that fish in that region have an unacceptably high mercury content.

a. Assuming that a mercury concentration of 5 ppm is considered the maximum safe concentration, which of the following pairs of hypotheses would you test:

$$H_0: \mu = 5 \quad \text{versus} \quad H_a: \mu > 5$$

or

$$H_0: \mu = 5 \quad \text{versus} \quad H_a: \mu < 5$$

Give the reasons for your choice.

b. Would you prefer a significance level of 0.10 or 0.01 for your test? Explain. (Hint: See discussion just before Example 10.7.)

10.21 The paper **"Living Near Nuclear Power Plants and Thyroid Cancer Risks" (*Environmental International* [2016]: 42–48)** investigated whether living near a nuclear power plant increases the risk of thyroid cancer. The authors of this paper concluded that there was no evidence of increased risk of thyroid cancer in areas that were near a nuclear power plant.

a. Let p denote the proportion of the population in areas near nuclear power plants who are diagnosed with thyroid cancer during a given year. The researchers who wrote this paper might have considered the two rival hypotheses of the form

H_0: p is equal to the corresponding value for areas without nuclear power plants
H_a: p is greater than the corresponding value for areas without nuclear power plants

Did the researchers reject H_0 or fail to reject H_0?

b. If the researchers are incorrect in their conclusion that there is no evidence of increased risk of thyroid cancer associated with living near a nuclear power plant, are they making a Type I or a Type II error? Explain.

c. Can the result of this hypothesis test be interpreted as meaning that there is strong evidence that the risk of thyroid cancer is not higher for people living near nuclear power plants? Explain.

10.22 An automobile manufacturer is considering using robots for part of its assembly process. Converting to robots is an expensive process, so it will be undertaken only if there is strong evidence that the proportion of defective installations is lower for the robots than for human assemblers. Let p denote the proportion of defective installations for the robots. It is known that human assemblers have a defect proportion of 0.02.

a. Which of the following pairs of hypotheses should the manufacturer test:

$$H_0\text{: } p = 0.02 \quad \text{versus} \quad H_a\text{: } p < 0.02$$

or

$$H_0\text{: } p = 0.02 \quad \text{versus} \quad H_a\text{: } p > 0.02$$

Explain your answer.

b. In the context of this exercise, describe Type I and Type II errors.

c. Would you prefer a test with $\alpha = 0.01$ or $\alpha = 0.10$? Explain your reasoning.

Section 10.3 Large-Sample Hypothesis Tests for a Population Proportion

Now that some basic concepts have been introduced, you are ready to consider how to use sample data to decide between a null and an alternative hypothesis. In a hypothesis test, there are two possible conclusions: You either reject H_0 or you fail to reject H_0. The logic of hypothesis-testing procedures is this: You reject the null hypothesis if the observed sample is very unlikely to have occurred when H_0 is true.

This section considers testing hypotheses about a population proportion when the sample size n is large. As before, p denotes the proportion of individuals or objects in a specified population that possess a certain property. A random sample of n individuals is selected from the population. The sample proportion

$$\hat{p} = \frac{\text{number in the sample that possess the property}}{n}$$

serves as the basis for testing hypotheses about p.

The large-sample test procedure is based on the same properties of the sampling distribution of $\hat{p}$ that were used to obtain a confidence interval for p:

1. $\mu_{\hat{p}} = p$
2. $\sigma_{\hat{p}} = \sqrt{\dfrac{p(1-p)}{n}}$
3. When n is large, the sampling distribution of $\hat{p}$ is approximately normal.

These three results imply that the standardized variable

$$z = \frac{\hat{p} - p}{\sqrt{\dfrac{p(1-p)}{n}}}$$

has approximately a standard normal distribution when n is large. Example 10.8 shows how this information allows you to make a decision in a hypothesis test.

Example 10.8 Impact of Food Labels

Understand the context ➤

An **Associated Press** survey was conducted to investigate how people use the nutritional information provided on food package labels (***Spartansburg Herald*, March 19, 2016**). Interviews were conducted with 1003 randomly selected adult Americans, and each participant was asked a series of questions, including the following two:

Question 1: When purchasing packaged food, how often do you check the nutrition labeling on the package?

Question 2: How often do you purchase foods that are bad for you, even after you've checked the nutrition labels?

Copyright 2025 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s). Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

It was reported that 582 responded "frequently" to the question about checking labels and 441 responded very often or somewhat often to the question about purchasing "bad" foods even after checking the label.

Start by considering the responses to the first question. Based on these data, is it reasonable to conclude that a majority of adult Americans frequently check the nutritional labels when purchasing packaged foods? You can answer this question by considering

p = actual proportion of all adult Americans who frequently check nutritional labels

and testing the following hypotheses:

H_0: $p = 0.5$
H_a: $p > 0.5$ (The proportion of adult Americans who frequently check nutritional labels is greater than 0.5. That is, more than half [a majority] frequently check nutritional labels.)

Recall that in a hypothesis test, the null hypothesis is rejected only if there is convincing evidence against it—in this case, convincing evidence that $p > 0.5$. If H_0 is rejected, there is strong support for the claim that a majority of adult Americans frequently check nutritional labels when purchasing packaged foods.

For this sample,

$$\hat{p} = \frac{582}{1003} = 0.58$$

The observed sample proportion is certainly greater than 0.5, but this could just be due to sampling variability. That is, when $p = 0.5$ (meaning H_0 is true), the sample proportion $\hat{p}$ usually differs somewhat from 0.5 simply because of chance variation from one sample to another. Is it plausible that a sample proportion of $\hat{p} = 0.58$ occurred because of this chance variation, or would it be unusual to observe a sample proportion this large when $p = 0.5$?

This question is answered by forming a *test statistic,* which is then used as a basis for making a decision between H_0 and H_a. Creating a test statistic involves replacing p with the hypothesized value in $z = \frac{\hat{p} - p}{\sqrt{\frac{p(1-p)}{n}}}$ to obtain

$$z = \frac{\hat{p} - 0.5}{\sqrt{\frac{(0.5)(1-0.5)}{n}}}$$

If the null hypothesis is true and the sample size is large, this statistic should have approximately a standard normal distribution, because in this case

1. $\mu_{\hat{p}} = 0.5$
2. $\sigma_{\hat{p}} = \sqrt{\frac{(0.5)(1-0.5)}{n}}$
3. $\hat{p}$ has approximately a normal distribution.

The calculated value of z is the difference between $\hat{p}$ and the hypothesized value of p in terms of the standard deviation. For example, if $z = 3$, then the value of the sample proportion $\hat{p}$ is 3 standard deviations (of $\hat{p}$) greater than what would be expected if the null hypothesis were true. How likely is it that a z value at least this inconsistent with H_0 would be observed if H_0 is true? If H_0 is true, the test statistic has (approximately) a standard normal distribution. This means that

$$P(z \geq 3 \text{ when } H_0 \text{ is true}) = \text{area under the } z \text{ curve to the right of } 3.00 = 0.0013$$

It follows that if H_0 is true, fewer than 1% of all samples would result in a value of z at least as inconsistent with H_0 as $z = 3$. Because this z value is in the most extreme 1% of the z distribution, it is reasonable to reject H_0.

For the given sample data,

Do the work ➤

$$z = \frac{\hat{p} - 0.5}{\sqrt{\frac{(0.5)(1 - 0.5)}{n}}} = \frac{0.58 - 0.5}{\sqrt{\frac{(0.5)(0.5)}{1003}}} = \frac{0.08}{0.016} = 5.00$$

This means that $\hat{p} = 0.58$ is 5 standard deviations greater than what would be expected if the null hypothesis H_0: $p = 0.5$ was true. The sample data appear to be much more consistent with the alternative hypothesis, H_a: $p > 0.5$. In particular,

$$\begin{aligned} &P(\text{value of } z \text{ is at least as contradictory to } H_0 \text{ as } 5.00 \text{ when } H_0 \text{ is true}) \\ &\quad = P(z \geq 5.00 \text{ when } H_0 \text{ is true}) \\ &\quad = \text{area under the } z \text{ curve to the right of } 5.00 \\ &\quad \approx 0 \end{aligned}$$

Interpret the results ➤ When the null hypothesis is true, there is virtually no chance of seeing a sample proportion and corresponding z value this extreme as a result of chance variation alone. If $\hat{p}$ is 5 standard deviations or more away from 0.5, it is hard to believe that $p = 0.5$. The evidence for rejecting H_0 in favor of H_a is very compelling.

In spite of the fact that there is strong evidence that a majority of adult Americans frequently check nutritional labels, the responses to the second question suggest that the percentage of people who then ignore the information on the label and purchase "bad" foods anyway is not small—the sample proportion who responded very often or somewhat often was 0.44.

The preceding example illustrates the reasoning behind large-sample procedures for testing hypotheses about p (and other test procedures as well). You begin by assuming that the null hypothesis is true. The sample is then examined in light of this assumption. If the observed sample proportion would not be unusual when H_0 is true, then chance variability from one sample to another is a plausible explanation for what has been observed, and H_0 is not rejected. On the other hand, if the observed sample proportion would have been unlikely when H_0 is true, the sample provides evidence against the null hypothesis and H_0 is rejected. The decision to reject or to fail to reject the null hypothesis is based on an assessment of how extreme or unlikely the observed sample is if H_0 is true.

The assessment of how inconsistent the observed sample proportion is with the null hypothesis, H_0, is based on the value of the test statistic

$$z = \frac{\hat{p} - \text{hypothesized value}}{\sqrt{\frac{(\text{hypothesized value})(1 - \text{hypothesized value})}{n}}}$$

The value of the test statistic is then used to calculate the ***P*-value**, the probability, *assuming that H_0 is true*, of obtaining a z value at least as inconsistent with H_0 as what was actually observed.

Definitions

Test statistic: A value calculated using sample data. It is the value used to make the decision between rejecting and failing to reject H_0.

***P*-value:** A measure of inconsistency between the hypothesized value for a population characteristic and the observed sample. It is the probability, *assuming that H_0 is true*, of obtaining a test statistic value at least as inconsistent with H_0 as what was observed. The *P*-value is also sometimes called the **observed significance level**.

Example 10.9 ChatGPT in Schools

Understand the context ➤ The introduction of artificial intelligence (AI) tools such as ChatGPT has led to concern about how students are using AI and whether it poses a threat to academic integrity. Researchers at study.com surveyed students and educators to gain insight into the use of AI in schools. The online article **"Productive Teaching Tool or Innovative Cheating" (study.com, January, 2023)** states that "Over one third of all educators believe that ChatGPT should be banned in schools and universities." This statement was based on a survey of 100 educators, with 34 responding that they believed ChatGPT should be banned.

Assuming it is reasonable to regard the sample as representative of educators, does the sample provide convincing evidence in support of the claim that more than one-third of all educators believe that ChatGPT should be banned in schools and universities

With

p = proportion of all educators who believe that ChatGPT should be banned in schools and universities

the relevant hypotheses are

$$H_0\colon p = \frac{1}{3} = 0.33$$
$$H_a\colon p > 0.33$$

The sample proportion is $\hat{p} = \dfrac{34}{100} = 0.34$.

Formulate a plan ➤ Is the value of $\hat{p}$ enough greater than one-third that you should reject H_0?

Because the sample size is large, the statistic

$$z = \frac{\hat{p} - 0.33}{\sqrt{\dfrac{(0.33)(1 - 0.33)}{n}}}$$

has approximately a standard normal distribution when H_0 is true. The calculated value of the test statistic is

Do the work ➤

$$z = \frac{0.34 - 0.33}{\sqrt{\dfrac{(0.33)(1 - 0.33)}{100}}} = \frac{0.01}{0.05} = 0.20$$

The probability that a z value at least this inconsistent with H_0 would be observed if H_0 is true is

$$\begin{aligned} P\text{-value} &= P(z \geq 0.20 \text{ when } H_0 \text{ is true}) \\ &= \text{area under the } z \text{ curve to the right of } 0.20 \\ &= 1 - 0.5793 \\ &= 0.4207 \end{aligned}$$

Technology could also have been used to calculate this *P*-value.

Interpret the results ➤ This probability indicates that when $p = 0.33$, it would not be unusual to observe a sample proportion as large as 0.34. When H_0 is true, about 42% of all samples would have a sample proportion as large as or larger than 0.34. This means that a sample proportion of 0.34 is reasonably consistent with the null hypothesis. Although 0.34 is larger than the hypothesized value of $p = 0.33$, chance variation from sample to sample is a plausible explanation for what was observed. There is not convincing evidence that the proportion of all educators who believe that ChatGPT should be banned in schools and universities is greater than one-third.

As illustrated by Examples 10.8 and 10.9, small *P*-values indicate that sample results are inconsistent with H_0, whereas larger *P*-values are interpreted as meaning that the data are consistent with H_0 and that sampling variability alone is a plausible explanation for what was observed in the sample. As you probably noticed, the two cases examined (P-value ≈ 0 and P-value $= 0.4207$) were such that a decision between rejecting or not rejecting H_0 was clear-cut. A decision in other cases might not be so obvious. For example, what if the sample had resulted in a P-value of 0.04? Is this unusual enough that H_0 should be rejected? How small does the P-value need to be before H_0 should be rejected?

The answers to these questions depend on the significance level, α (the probability of a Type I error), selected for the test. For example, suppose that you set $\alpha = 0.05$. This implies that the probability of rejecting a true null hypothesis is 0.05. To obtain a test procedure with this probability of Type I error, you would reject the null hypothesis if the sample result is among the most unusual 5% of all samples when H_0 is true. You would reject H_0 if the calculated P-value ≤ 0.05. If you had selected $\alpha = 0.01$, H_0 would be rejected only if you observed a sample result so extreme that it would be among the most unusual 1% if H_0 is true (which occurs when P-value ≤ 0.01).

A decision to reject or to fail to reject H_0 results from comparing the P-value to the chosen significance level α:

Reject H_0 if P-value $\leq \alpha$.

Fail to reject H_0 if P-value $> \alpha$.

Suppose, for example, that the P-value $= 0.0352$ and that a significance level of 0.05 is chosen. Then, because

$$P\text{-value} = 0.0352 \leq 0.05 = \alpha$$

H_0 would be rejected. This would not be the case, though, for $\alpha = 0.01$, because then P-value $> \alpha$.

Calculating a *p*-Value for a Large-Sample Test About *p*

The calculation of the P-value depends on the form of the inequality in the alternative hypothesis, H_a. For example, suppose that you want to test

$$H_0\colon\ p = 0.6 \qquad \text{versus} \qquad H_a\colon\ p > 0.6$$

based on a large sample. The appropriate test statistic is

$$z = \frac{\hat{p} - 0.6}{\sqrt{\dfrac{(0.6)(1 - 0.6)}{n}}}$$

Values of $\hat{p}$ inconsistent with H_0 and much more consistent with H_a are those much *greater* than 0.6 (because $p = 0.6$ when H_0 is true and $p > 0.6$ when H_0 is false and H_a is true). These values of $\hat{p}$ correspond to z values considerably greater than 0.

If $n = 400$ and $\hat{p} = 0.679$, then

$$z = \frac{0.679 - 0.6}{\sqrt{\dfrac{(0.6)(1 - 0.6)}{400}}} = \frac{0.079}{0.024} = 3.29$$

The value $\hat{p} = 0.679$ is more than 3 standard deviations larger than what would have been expected if H_0 were true. Then,

$$\begin{aligned} P\text{-value} &= P(z \text{ at least as inconsistent with } H_0 \text{ as } 3.29 \text{ when } H_0 \text{ is true}) \\ &= P(z \geq 3.29) \end{aligned}$$

$$\begin{aligned} &= \text{area under the } z \text{ curve to the right of } 3.29 \\ &= 1 - 0.9995 \\ &= 0.0005 \end{aligned}$$

This P-value is illustrated in Figure 10.1. If H_0 is true, in the long run, only 5 out of 10,000 samples would result in a z value as or more extreme than what actually resulted. Most people would consider this quite unusual. Using a significance level of 0.01, you would reject the null hypothesis because $P\text{-value} = 0.0005 \leq 0.01 = \alpha$.

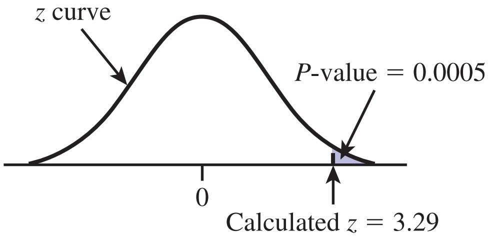


Figure 10.1
Calculating a P-value.

What Do You Think?

How do you think the procedure for finding a P-value would be different if the test is a two-sided test?

Now consider testing H_0: $p = 0.3$ versus H_a: $p \neq 0.3$. A value of $\hat{p}$ *either* much greater than 0.3 *or* much less than 0.3 is inconsistent with H_0 and provides support for H_a. These $\hat{p}$ values correspond to a z value far out in *either* tail of the z curve. If

$$z = \frac{\hat{p} - 0.3}{\sqrt{\frac{(0.3)(1 - 0.3)}{n}}} = 1.75$$

then (as shown in Figure 10.2)

$$\begin{aligned} P\text{-value} &= P(z \text{ value at least as inconsistent with } H_0 \text{ as } 1.75 \text{ when } H_0 \text{ is true}) \\ &= P(z \geq 1.75 \text{ or } z \leq -1.75) \\ &= (z \text{ curve area to the right of } 1.75) + (z \text{ curve area to the left of } -1.75) \\ &= (1 - 0.9599) + 0.0401 \\ &= 0.0802 \end{aligned}$$

If $z = -1.75$, the P-value in this situation is also 0.0802 because 1.75 and -1.75 are equally inconsistent with H_0.

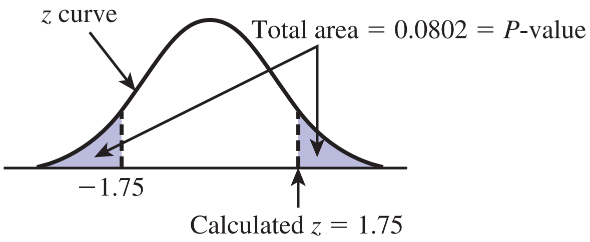


Figure 10.2
P-value as the sum of two tail areas.

The symmetry of the z curve implies that when the test is two-tailed (the "not equal" alternative), it is not necessary to add two curve areas. Instead,

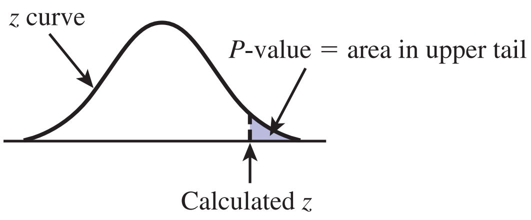
If z is positive, $P\text{-value} = 2(\text{area to the right of } z)$.
If z is negative, $P\text{-value} = 2(\text{area to the left of } z)$.

Determination of the *p*-Value When the Test Statistic Is *z*

1. **Upper-tailed test:**

   H_a: $p >$ hypothesized value

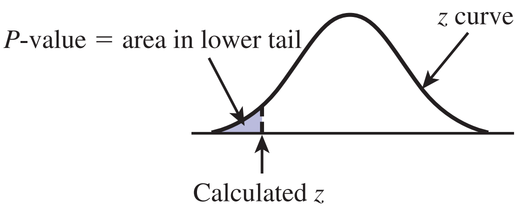
   P-value calculated as illustrated:



2. **Lower-tailed test:**

   H_a: $p <$ hypothesized value

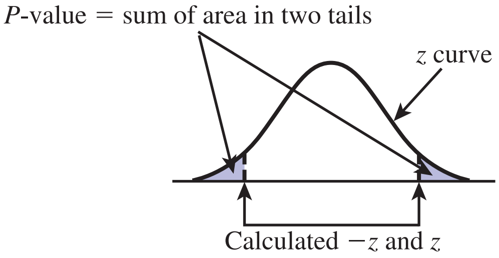
   P-value calculated as illustrated:



3. **Two-tailed test:**

   H_a: $p \neq$ hypothesized value

   P-value calculated as illustrated:



A statistical software package or a graphing calculator can also be used to carry out the large-sample z test for a population proportion. This is illustrated in Examples 10.10 and 10.11.

Example 10.10 Water Conservation

Understand the context ➤

Suppose that in December 2022, a county-wide water conservation campaign was conducted in a particular county. In January 2023, a random sample of 500 homes is selected, and water usage is recorded for each home in the sample. The county supervisors wanted to know whether the data supported the claim that less than half of the households in the county reduced water consumption. The relevant hypotheses are

$$H_0\colon\ p = 0.5 \qquad \text{versus} \qquad H_a\colon\ p < 0.5$$

where p is the proportion of all households in the county with reduced water usage.

Formulate a plan ➤

Suppose that the sample results were $n = 500$ and $\hat{p} = 0.440$. Because the sample size is large and this is a lower-tailed test, you can calculate the P-value by first calculating the value of the z test statistic

$$z = \frac{\hat{p} - 0.5}{\sqrt{\dfrac{(0.5)(1 - 0.5)}{n}}}$$

Copyright 2025 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s). Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

and then finding the area under the z curve to the left of this z.

Based on the observed sample data,

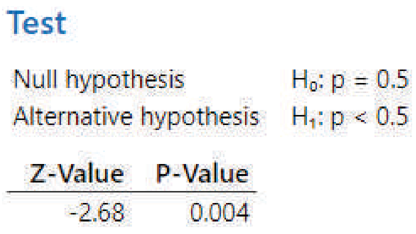
Do the work ➤

$$z = \frac{0.440 - 0.5}{\sqrt{\frac{(0.5)(1-0.5)}{500}}} = \frac{-0.060}{0.0224} = -2.68$$

The P-value is then equal to the area under the z curve and to the left of -2.68. From the entry in the -2.6 row and 0.08 column of Appendix Table 2,

P-value $= 0.0037$

Technology could also have been used to find the value of the test statistic and the P-value. Using Minitab produces the following output:



```
Test

Null hypothesis          H₀: p = 0.5
Alternative hypothesis   H₁: p < 0.5

 Z-Value   P-Value
 -2.68     0.004
```

Notice that the test statistic and P-value in the Minitab output agree with the hand calculations, although the P-value in Minitab has been rounded to three decimal places.

Interpret the results ➤

Using a 0.01 significance level, you would reject H_0 (because $0.0037 \leq 0.01$). This leads to a conclusion that there is convincing evidence that the proportion with reduced water usage was less than 0.5. Notice that H_0 would not have been rejected if a *very* small significance level, such as 0.001, had been selected.

Example 10.10 illustrates the calculation of a P-value for a lower-tailed test. The use of P-values in upper-tailed and two-tailed tests will be illustrated in Examples 10.11 and 10.12. But first the large-sample hypothesis test for a population proportion is summarized and a step-by-step procedure for carrying out a hypothesis test is introduced.

Large-Sample *z* Test for a population proportion *p*

Null Hypothesis: H_0: p = hypothesized value

Test Statistic:

$$z = \frac{\hat{p} - \text{hypothesized value}}{\sqrt{\frac{(\text{hypothesized value})(1 - \text{hypothesized value})}{n}}}$$

Alternative Hypothesis:	***P*-Value:**
H_a: $p >$ hypothesized value	Area under z curve to right of calculated z
H_a: $p <$ hypothesized value	Area under z curve to left of calculated z
H_a: $p \neq$ hypothesized value	(1) 2(area to right of z) if z is positive, or (2) 2(area to left of z) if z is negative

Assumptions:

1. $\hat{p}$ is the sample proportion from a random sample.
2. The sample size is large. This test can be used if n satisfies both n(hypothesized value) ≥ 10 and $n(1 -$ hypothesized value$) \geq 10$.
3. If sampling is without replacement, the sample size is no more than 10% of the population size.

Copyright 2025 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s). Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

The following sequence of steps can be used when carrying out a hypothesis test.

Steps in a Hypothesis Test

1. Describe the population characteristic of interest.
2. State the null hypothesis H_0.
3. State the alternative hypothesis H_a.
4. Select the significance level α for the test.
5. Display the test statistic to be used, with substitution of the hypothesized value identified in Step 2.
6. Check to make sure that any assumptions required for the test are reasonable.
7. Calculate all quantities appearing in the test statistic and the value of the test statistic itself.
8. Determine the *P*-value associated with the observed value of the test statistic.
9. State the conclusion (which is reject H_0 if $P\text{-value} \le \alpha$ and fail to reject H_0 otherwise). The conclusion should then be stated in the context of the problem. The level of significance should be included when stating the conclusion.

Steps 1–4 constitute a statement of the problem, Steps 5–8 give the analysis that leads to a decision, and Step 9 provides the conclusion.

Example 10.11 Love Those Cell Phones . . .

Understand the context ➤

The article **"You Now Have a Shorter Attention Span Than a Goldfish" (*Time*, May 14, 2015)** describes a study of 2000 Canadians over the age of 18 that was carried out by Microsoft. Study participants were asked whether the following statement described them: "When nothing is occupying my attention, the first thing I do is reach for my phone." Of the study participants in the age group 18 to 24 years old, 77% responded "yes" to this question. Suppose that this group of 18- to 24-year-olds can be considered as a representative sample of Canadians in this age group, and suppose that 800 of the study participants were in this age group.

Does this sample support the claim that more than 75% of all Canadians in this age group would respond "yes" to the given statement?

You can answer this question by following the nine steps for carrying out a hypothesis test. A 0.05 significance level will be used for this example.

1. Population characteristic of interest:

   p = proportion of all Canadians age 18 to 24 who would say that the given statement describes them

2. Null hypothesis: H_0: $p = 0.75$
3. Alternative hypothesis: H_a: $p > 0.75$ (the percentage of Canadians age 18 to 24 who would say that the given statement describes them is greater than 75%)
4. Significance level: $\alpha = 0.05$

Formulate a plan ➤

5. Test statistic:

$$z = \frac{\hat{p} - \text{hypothesized value}}{\sqrt{\dfrac{(\text{hypothesized value})(1 - \text{hypothesized value})}{n}}} = \frac{\hat{p} - 0.75}{\sqrt{\dfrac{(0.75)(1 - 0.75)}{n}}}$$

6. Assumptions: This test requires a random sample and a large sample size. The given sample was considered to be representative of Canadians age 18 to 24, and if this is the case, it is reasonable to regard the sample as if it were a random sample. The sample size was $n = 800$. Since $800(0.75) = 600 \ge 10$ and $800\,(1 - 0.75) = 200 \ge 10$, the large-sample test is appropriate. The population is all Canadians age 18 to 24, so the sample size is small compared to the population size.

Copyright 2025 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s). Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

Do the work ➤

7. Calculations: $n = 800$ and $\hat{p} = 0.77$, so

$$z = \frac{\hat{p} - 0.75}{\sqrt{\frac{0.75(1 - 0.75)}{800}}} = \frac{0.77 - 0.75}{\sqrt{\frac{0.75(0.25)}{800}}} = \frac{0.02}{0.015} = 1.33$$

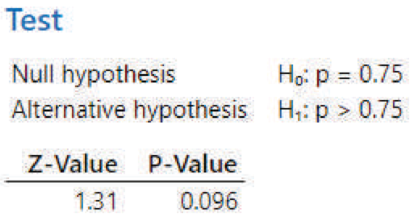
8. *P*-value: This is an upper-tailed test (the inequality in the null hypothesis is $>$), so the *P*-value is the area under the z curve and to the right of the calculated z value.

$$\begin{aligned} P\text{-value} &= \text{area to the right of } 1.33 \\ &= P(z > 1.33) \\ &= 0.0918 \end{aligned}$$

Interpret the results ➤

9. Conclusion: Because the *P*-value is greater than the selected significance level, you fail to reject the null hypothesis. You can conclude that the sample does not provide convincing evidence that more than 75% of Canadians age 18 to 24 think that the statement "When nothing is occupying my attention, the first thing I do is reach for my phone" describes them.

Technology could also have been used to complete steps 7 and 8. The small differences in the value of the test statistic and the *P*-value are just due to rounding in the hand calculations.



Test

Null hypothesis	H_0: p = 0.75
Alternative hypothesis	H_1: p > 0.75

Z-Value	P-Value
1.31	0.096

Example 10.12 College Attendance

Understand the context ➤

The report **"2019–2020 College-Going Rate for California High School Students" (California Department of Education, cde.ca.gov)** indicates that 63% of students graduating from California high schools go on to attend a two-year or four-year college the year after graduation. This college-going rate for students graduating from high schools in the Los Angeles Unified School District was estimated to be 52%. Suppose that the estimate of 52% was based on a representative sample of 500 graduates from high schools in the Los Angeles Unified School District.

Is it reasonable to conclude that the proportion of Los Angeles Unified School District high school graduates who attended college the year after graduation is different from the statewide proportion of 0.63? The nine-step hypothesis testing procedure and a significance level of $\alpha = 0.01$ will be used to answer this question.

1. p = proportion of all Los Angeles Unified School District high school graduates who attend college the year after graduation

2. H_0: $p = 0.63$

3. H_a: $p \neq 0.63$ (differs from the statewide proportion)

4. Significance level: $\alpha = 0.01$

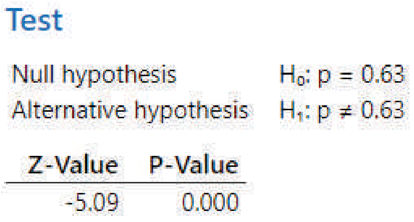
Formulate a plan ➤

5. Test statistic:

$$z = \frac{\hat{p} - \text{hypothesized value}}{\sqrt{\frac{(\text{hypothesized value})(1 - \text{hypothesized value})}{n}}} = \frac{\hat{p} - 0.63}{\sqrt{\frac{(0.63)(1 - 0.63)}{n}}}$$

6. Assumptions: This test requires a random sample and a large sample size. The sample was said to be representative of Los Angeles Unified School District high school graduates, so it is reasonable to regard the sample as if it were a random sample. The sample size was $n = 500$, and the population size is much larger than the sample size. Because $500(0.63) \geq 10$ and $500(1 - 0.63) \geq 10$, the large-sample test is appropriate.

Do the work ➤ **7.** Using technology to calculate the test statistic and P-value results in the following output:



```
Test

Null hypothesis          H₀: p = 0.63
Alternative hypothesis   H₁: p ≠ 0.63

Z-Value   P-Value
  -5.09     0.000
```

From the output, $z = -5.09$.

8. From the output given instep 7, P-value $= 0.000$

Interpret the results ➤ **9.** Conclusion: At significance level 0.01, you would reject H_0 because the P-value, rounded to three decimal places, is 0.000, which is less than the significance level of $\alpha = 0.01$. The data provides convincing evidence that the actual proportion of Los Angeles Unified School District high school graduates who attended college during the year after graduation differs from the statewide proportion.

Exercises 10.23–10.42

10.23 Use the definition of the P-value to explain the following:

a. Why H_0 would be rejected if P-value $= 0.0003$

b. Why H_0 would not be rejected if P-value $= 0.350$

10.24 For which of the following P-values will the null hypothesis be rejected when performing a test with a significance level of 0.05:

a. 0.001

b. 0.021

c. 0.078

d. 0.047

e. 0.148

10.25 Pairs of P-values and significance levels, α, are given. For each pair, state whether you would reject H_0 at the given significance level.

a. P-value $= 0.084$, $\alpha = 0.05$

b. P-value $= 0.003$, $\alpha = 0.001$

c. P-value $= 0.498$, $\alpha = 0.05$

10.26 Pairs of P-values and significance levels are given. For each pair, state whether you would reject the null hypothesis at the given significance level.

a. P-value $= 0.084$, $\alpha = 0.10$

b. P-value $= 0.039$, $\alpha = 0.01$

c. P-value $= 0.218$, $\alpha = 0.10$

10.27 Let p denote the proportion of students at a particular university that use the fitness center on campus on a regular basis. For a large-sample z test of H_0: $p = 0.5$ versus H_a: $p > 0.5$, find the P-value associated with each of the given values of the test statistic:

a. 1.40

b. 0.93

c. 1.96

d. 2.45

e. -0.17

10.28 Assuming a random sample from a large population, for which of the following null hypotheses and sample sizes n is the large-sample z test appropriate:

a. H_0: $p = 0.2$, $n = 25$

b. H_0: $p = 0.6$, $n = 210$

c. H_0: $p = 0.9$, $n = 100$

d. H_0: $p = 0.05$, $n = 75$

10.29 The article **"The World Prefers a Calm Life to an Exciting Life" (gallup.com, September 30, 2021)** reports the results of a survey by the Wellbeing for Planet Earth (WPE) Foundation that asked adults from 116 countries and territories around the world, "Would you rather live an exciting life or a calm life?" The resulting data were used to test H_0: $p = 0.5$ versus H_a: $p > 0.5$, and the null hypothesis was rejected.

a. Based on the title of the article, what do you think p represents in the hypotheses? (Hint: Make sure to distinguish between a sample statistic and a population parameter.)

b. Is it reasonable to say that the data provide strong support for the alternative hypothesis?

Copyright 2025 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s). Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

c. Is it reasonable to say that the data provide strong evidence against the null hypothesis?

10.30 In 2023, Major League Baseball (MLB) implemented several rule changes designed to shorten game length. In a survey of a representative sample of 347 baseball fans conducted by YouGov **("How Major League Baseball Fans View the New Rules of the Game," yougov.com, March 29, 2023)**, 194 said that they support the new pitch clock rule. This new rule gives pitchers 15 seconds to throw a pitch when the bases are empty and 20 seconds when a runner is on base.

a. Based on this sample, is there convincing evidence that a majority of MLB fans support the new pitch clock rule? Test the relevant hypotheses using a significance level of 0.05.

b. Would you reach the same conclusion if a significance level of 0.01 had been used for the test?

10.31 The survey referenced in the previous exercise also asked the people in the representative sample of baseball fans if they supported a new rule that increased the size of the bases from 15 inches to 18 inches. This rule was designed to reduce collisions and to encourage more stolen-base attempts. Of the 347 baseball fans surveyed, 170 said that they support this new rule.

a. Based on this sample, is there convincing evidence that fewer than half of MLB fans support the base size rule? Test the relevant hypotheses using a significance level of 0.05.

b. Would you reach the same conclusion if a significance level of 0.01 had been used for the test?

10.32 The paper **"Debt Literacy, Financial Experiences and Over-Indebtedness" (*Social Science Research Network,* Working paper W14808, 2008)** included analysis of data from a national sample of 1000 Americans. One question on the survey was:

> "You owe \$3000 on your credit card. You pay a minimum payment of \$30 each month. At an Annual Percentage Rate of 12% (or 1% per month), how many years would it take to eliminate your credit card debt if you made no additional charges?"

Answer options for this question were: (a) less than 5 years; (b) between 5 and 10 years; (c) between 10 and 15 years; (d) never—you will continue to be in debt; (e) don't know; and (f) prefer not to answer.

a. Only 354 of the 1000 respondents chose the correct answer of never. For purposes of this exercise, assume that the sample is representative of adult Americans. Is there convincing evidence that the proportion of adult Americans who can answer this question correctly is less than 0.40 (40%)? Use $\alpha = 0.05$ to test the appropriate hypotheses. (Hint: See Example 10.10.)

b. The paper also reported that 37.8% of those in the sample chose one of the wrong answers (a, b, and c) as their response to this question. Is it reasonable to conclude that more than one-third of adult Americans would select a wrong answer to this question? Use $\alpha = 0.05$.

10.33 The report **"Survey Finds Half of Consumers Connect to the Internet in Their Cars," (aftermarketnews.com, June 28, 2018)** summarized results from a survey of 1,200 car owners. Of those surveyed, 80% indicated that they are happy with the technology currently available in their cars. Assuming that it is reasonable to consider this sample as a representative sample of car owners, do the sample data provide convincing evidence than more than ¾ of all car owners are satisfied with the technology currently available in their cars? Test the relevant hypotheses using a significance level of 0.05

10.34 The article **"Euthanasia Still Acceptable to Solid Majority in U.S." (gallup.com, June 24, 2016)** summarized data from a survey of 1025 adult Americans. When asked if doctors should be able to end a terminally ill patient's life by painless means if requested to do so by the patient, 707 of those surveyed responded yes. For purposes of this exercise, assume that it is reasonable to regard this sample as a random sample of adult Americans. Suppose that you want to use the data from this survey to decide if there is convincing evidence that more than two-thirds of adult Americans believe that doctors should be able to end a terminally ill patient's life if requested to do so by the patient.

a. What hypotheses should be tested to answer this question?

b. The *P*-value for this test is 0.058. What conclusions would you reach if $\alpha = 0.05$?

c. Would you have reached a different conclusion if $\alpha = 0.10$? Explain.

10.35 The report **"Digital Democracy Survey" (Deloitte Development LLC, 2016, deloitte.com)** says that 69% of U.S. teens access social media from a mobile phone. Because this is based on data from 2016, you plan to select a random sample of students at the local high school and will ask each student in the sample if they access social media from a mobile phone. You want to determine if there is evidence that the proportion of students at the high school who access social media using a mobile phone differs from the 2016 national figure of 0.69 given in the Nielsen report. What hypotheses should you test?

10.36 According to a data intelligence company that specializes in the short-term rental market, 79% of Airbnb listings are for entire homes or apartments, as opposed to shared spaces **(seetransparent.com, November 4, 2020)**. Do Airbnb listings in Seattle have the same distribution of housing types? Using publicly available data, suppose you take a random sample of 100 Airbnb listings in Seattle. Of those, 84 were for entire homes or apartments. Is there sufficient evidence to conclude that the percentage of listings in Seattle that are for entire homes or apartments is different from 79%? Use a significance level of $\alpha = 0.05$ to test the relevant hypotheses.

10.37 The article **"Facebook Use and Academic Performance Among College Students" (*Computers in Human Behavior* [2015]: 265–272)** estimated that 87% percent of students at a large public university in California who are Facebook users update their status at least two times a day. Suppose that you plan to select a random sample of 400 students at your college. You will ask each student in the sample if they are a Facebook user and if they update their status at least two times a day. You plan to use the resulting data to decide if there is evidence that the proportion for your college is different from the proportion reported in the article for the college in California. What hypotheses should you test?

10.38 The article **"Public Acceptability in the UK and the USA of Nudging to Reduce Obesity: The Example of Reducing Sugar-Sweetened Beverages" (*PLOS One*, June 8, 2016)** describes a survey in which each person in a representative sample of 1082 adult Americans was asked about whether they would find different types of interventions acceptable to reduce consumption of sugary beverages. When asked about a tax on sugary beverages, 459 of the people in the sample said they thought that this would be an acceptable intervention. These data were used to test H_0: $p = 0.5$ versus H_a: $p < 0.5$ and the null hypothesis was rejected. Based on the hypothesis test, what can you conclude about the proportion of adult Americans who think that taxing sugary beverages is an acceptable intervention to reduce consumption of sugary beverages?

10.39 The article titled **"7% of Americans Don't Use the Internet. Who Are They?"** describes a study conducted by the **Pew Research Center (April 2, 2021).** Suppose that the title of this article is based on a representative sample of 600 adult Americans. Does this support the claim that the proportion of all adult Americans who do not use the Internet is less than 0.10 (10%)?

10.40 The article **"The NFL's Overtime Rules Aren't Fair—But Neither Are the Alternatives" (*The Ringer*, February 6, 2017)** gives data for 82 National Football League (NFL) games that went to overtime between 2012 and 2017. In that sample of games, the team that won the coin flip at the start of overtime went on to win in 45 out of 82 games. The article claims that winning the coin toss is an advantage. Assuming that this sample of games is representative of NFL games that go into overtime, do these data provide convincing evidence that this claim is justified?

10.41 The article **"After Testing Four-Day Week, Companies Say They Don't Want to Stop" (*The Wall Street Journal*, February 21, 2023)** described a study in which 61 British businesses of various types and sizes gave their workers a paid day off each week for 6 months to see if they could get as much done by working more efficiently when they were at work. At the end of the 6 months, 18 of the companies planned to make the 4-day work week permanent.

a. What assumption would you need to make about the sample of businesses that participated in the study in order to use the large-sample z test for a population proportion to determine if there was sufficient evidence to conclude that more than 20% of all British businesses that implemented a 4-day work week would decide to make the change permanent.

b. Suppose that the assumption that you identified in Part (a) is reasonable. Test the relevant hypotheses.

10.42 The United States Elections Project **(electproject.org)** reported that 68.6% of registered voters in California voted in the 2020 presidential election and that this was greater than the national percentage of 66.6%. Explain why it is not necessary to carry out a hypothesis test to determine if the proportion of registered voters in California who voted in the election is greater than the national proportion of 0.666.

Section 10.4 Hypothesis Tests for a Population Mean

This section introduces a method for testing hypotheses about a population mean. The test procedures in this case are based on the same two results that led to the z and t confidence intervals in Chapter 9:

1. When n is large or the population distribution is approximately normal, then

$$z = \frac{\bar{x} - \mu}{\frac{\sigma}{\sqrt{n}}}$$

has approximately a standard normal distribution.

Copyright 2025 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s). Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

2. When n is large or the population distribution is approximately normal, then

$$t = \frac{\bar{x} - \mu}{\frac{s}{\sqrt{n}}}$$

has approximately a t distribution with df $= n - 1$.

This means that if you are interested in testing a null hypothesis of the form

H_0: μ = hypothesized value

and n is large or the population distribution is approximately normal, one of the following z or t test statistics can be used:

Case 1: σ known

Test statistic: $$z = \frac{\bar{x} - \text{hypothesized value}}{\frac{\sigma}{\sqrt{n}}}$$

P-value: Calculated as an area under the z curve

Case 2: σ unknown

Test statistic: $$t = \frac{\bar{x} - \text{hypothesized value}}{\frac{s}{\sqrt{n}}}$$

P-value: Calculated as an area under the t curve with df $= n - 1$

Because it is rarely the case that σ, the population standard deviation, is known, the focus here will be on the test procedure for the case in which σ is unknown.

When testing a hypothesis about a population mean, the null hypothesis specifies a particular hypothesized value for μ. The null hypothesis is written as H_0: μ = hypothesized value. The alternative hypothesis has one of the following three forms, depending on the research question being addressed:

H_a: $\mu >$ hypothesized value
H_a: $\mu <$ hypothesized value
H_a: $\mu \neq$ hypothesized value

If n is large or if the population distribution is approximately normal, the test statistic

$$t = \frac{\bar{x} - \text{hypothesized value}}{\frac{s}{\sqrt{n}}}$$

can be used. For example, if the null hypothesis is H_0: $\mu = 100$, the test statistic becomes

$$t = \frac{\bar{x} - 100}{\frac{s}{\sqrt{n}}}$$

Consider the alternative hypothesis H_a: $\mu > 100$, and suppose that a sample of size $n = 24$ results in $\bar{x} = 104.20$ and $s = 8.23$. Then the test statistic value is

$$t = \frac{104.20 - 100}{\frac{8.23}{\sqrt{24}}} = \frac{4.20}{1.6799} = 2.50$$

Because this is an upper-tailed test, the P-value is the area under an appropriate t curve (here with df $= 24 - 1 = 23$) to the right of 2.50.

If you have access to a statistics software package or a graphing calculator, this technology can be used to calculate the value of the *t* statistic and the *P*-value. If you are not using technology, Appendix Table 4 can be used to get approximate *P*-values. Appendix Table 4 gives *t* curve tail areas. Each column of the table is for a different number of degrees of freedom: 1, 2, 3, . . . , 30, 35, 40, 60, 120, and a last column for df = ∞, which is the same as for the *z* curve. The table gives the area under each *t* curve to the right of values ranging from 0.0 to 4.0 in increments of 0.1. Part of this table appears in Figure 10.3. For example,

$$\begin{aligned} \text{area under the 23-df } t \text{ curve to the right of } 2.5 &= 0.010 \\ &= P\text{-value for an upper-tailed } t \text{ test} \end{aligned}$$

Suppose that $t = -2.7$ for a lower-tailed test based on 23 df. Then, because each *t* curve is symmetric about 0,

$$P\text{-value} = \text{area to the left of } -2.7 = \text{area to the right of } 2.7 = 0.006$$

As is the case for *z* tests, you double the tail area to obtain the *P*-value for two-tailed *t* tests. This means that if $t = 2.6$ or if $t = -2.6$ for a two-tailed *t* test with 23 df, then

$$P\text{-value} = 2(0.008) = 0.016$$

Once past 30 df, the tail areas change very little, so the last column (∞) in Appendix Table 4 provides a good approximation.

Figure 10.3
Part of Appendix Table 4: *t* curve tail areas.

df *t*	1	2	. . .	22	23	24	. . .	60	120
0.0									
0.1									
⋮				⋮	⋮	⋮			
2.5			. . .	.010	.010	.010	. . .		
2.6			. . .	.008	.008	.008	. . .		
2.7			. . .	.007	.006	.006	. . .		
2.8			. . .	.005	.005	.005	. . .		
⋮				⋮	⋮	⋮			
4.0									

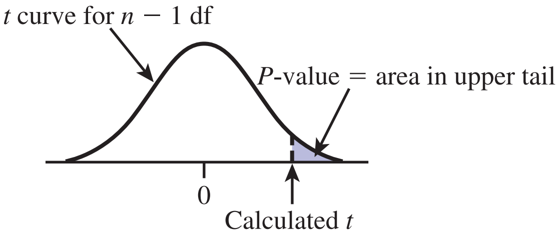
Area under 23-df t curve to right of 2.7

The following two boxes show how the *P*-value is obtained as a *t* curve area and give a general description of the test procedure.

Finding *p*-Values for a *t* Test

1. **Upper-tailed test:**

H_a: $\mu >$ hypothesized value

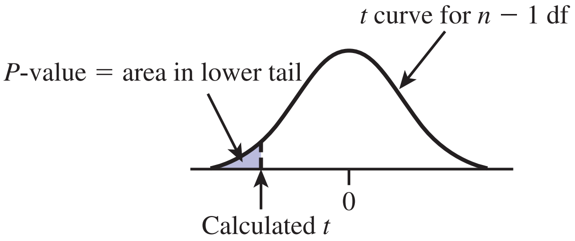


(continued)

Copyright 2025 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s). Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

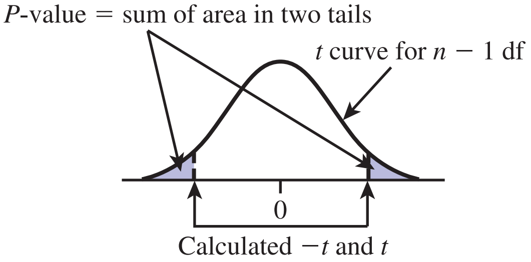
2. **Lower-tailed test:**

   H_a: $\mu <$ hypothesized value



3. **Two-tailed test:**

   H_a: $\mu \neq$ hypothesized value



P-values can be found using technology, or Appendix Table 4 gives upper-tail *t* curve areas to the right of values 0.0, 0.1, . . . , 4.0. These areas are *P*-values for upper-tailed tests and, by symmetry, also for lower-tailed tests. Doubling an area gives the *P*-value for a two-tailed test.

One-Sample *t* Test for a Population Mean

Null Hypothesis: H_0: $\mu =$ hypothesized value

Test Statistic: $t = \dfrac{\bar{x} - \text{hypothesized value}}{\dfrac{s}{\sqrt{n}}}$

Alternative Hypothesis:	***P*-Value:**
H_a: $\mu >$ hypothesized value	Area to the right of calculated t under t curve with df $= n - 1$
H_a: $\mu <$ hypothesized value	Area to the left of calculated t under t curve with df $= n - 1$
H_a: $\mu \neq$ hypothesized value	(1) 2(area to the right of t) if t is positive, or (2) 2(area to the left of t) if t is negative

Assumptions:
1. $\bar{x}$ and s are the sample mean and sample standard deviation from a random sample.
2. The sample size is large (generally $n \geq 30$) or the population distribution is at least approximately normal.

Example 10.13 Getting Enough Sleep

● The American Academy of Sleep Medicine recommends that teens 13 to 18 years of age should get 8 to 10 hours of sleep a night for optimal health. The authors of the paper **"Exposure to Electronic Devices and Sleep Quality in Adolescents: A Matter of Type, Duration, and Timing" (*Sleep Health* [2020]: 172–178)** describe a study in which students in this age group were asked about how much sleep they typically get on nights when they have school the next day. Data on reported sleep time (in minutes) for a representative sample of 10 students that are consistent with summary quantities given in the paper are shown below. Do these data provide evidence that the mean sleep time for students on nights when they have school the next day is less than the minimum recommendation of 8 hours (480 minutes)?

396 436 573 491 434 545 490 345 490 401

● **Data set available online**

Copyright 2025 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s). Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

The sample mean and standard deviation calculated from these data are

$$n = 10 \qquad \bar{x} = 460.1 \qquad s = 70.6$$

With μ representing the mean sleep time for students on a night when there is school the next day, you can answer this question by testing

$$H_0\!: \ \mu = 480 \text{ minutes}$$

versus

$$H_a\!: \ \mu < 480 \text{ minutes}$$

The null hypothesis is rejected only if there is convincing evidence that $\mu < 480$. The observed sample mean, 460.1, is less than 480, but can a sample mean as small as this be explained by chance variation from one sample to another when $\mu = 480$?

To answer this question, you can carry out a hypothesis test with a significance level of 0.05 using the nine-step procedure described in Section 10.3.

1. Population characteristic of interest:

Understand the context ➤

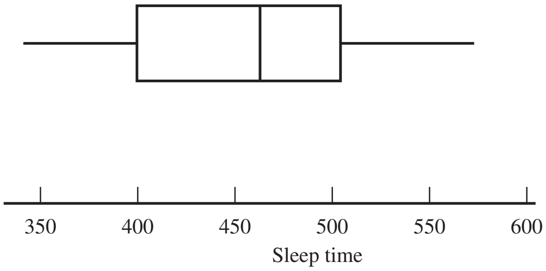
μ = mean sleep time for students on a night when there is school the next day

2. Null hypothesis: H_0: $\mu = 480$
3. Alternative hypothesis: H_a: $\mu < 480$
4. Significance level: $\alpha = 0.05$

Formulate a plan ➤

5. Test statistic: $t = \dfrac{\bar{x} - \text{hypothesized value}}{\frac{s}{\sqrt{n}}} = \dfrac{\bar{x} - 480}{\frac{s}{\sqrt{n}}}$

6. Assumptions: The one-sample t test requires a random sample and either a large sample size or a normal population distribution. The problem statement indicated that it was reasonable to consider this sample as representative of students in general, so it is reasonable to regard the sample as if it were a random sample. Because the sample size is only 10, for the t test to be appropriate, you must be willing to assume that the population distribution of sleep time is at least approximately normal. Is this reasonable? The following graph is a boxplot of the data:



Although the boxplot is not perfectly symmetric, it does not appear to be too skewed and there are no outliers, so the use of the t test is reasonable.

7. Calculations: $n = 10$, $\bar{x} = 460.1$, and $s = 70.6$, so

Do the work ➤

$$t = \frac{460.1 - 480}{\frac{70.6}{\sqrt{10}}} = \frac{-19.9}{22.33} = -0.89$$

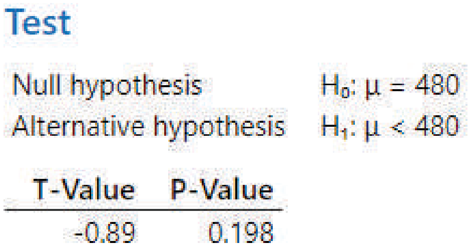
8. P-value: This is a lower-tailed test (the inequality in H_a is "less than"), so the P-value is the area to the left of the calculated t value. Because df $= 10 - 1 = 9$, you can use the df $= 9$ column of Appendix Table 4 to find the P-value. With $t = -0.89$, P-value = area to the left of -0.89. The area to the left of -0.89 is approximately 0.196 (using 0.9, which is the closest value to -0.89 that appears in the table). The P-value ≈ 0.196.

Copyright 2025 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s). Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

Interpret the results ➤

9. Conclusion: Because the *P*-value (0.196) is greater than the selected significance level, you fail to reject H_0 at the 0.05 level of significance. There is not convincing evidence that the population mean sleep time is less than the minimum recommendation of 8 hours (480 minutes). Even though the sample mean is less than 480 minutes, for a sample size of 10 it would not be surprising to see a sample mean as small as what was observed just due to sampling variability.

A statistical software package or a graphing calculator could also be used to carry out the calculate step of the hypothesis test process. For example, in the accompanying Minitab output, you can see that the value of the test statistic is given as $t = -0.89$ and the associated *P*-value is given as 0.198



Test

Null hypothesis H_0: μ = 480
Alternative hypothesis H_1: μ < 480

T-Value	P-Value
-0.89	0.198

The difference between what is given in the Minitab output and what was obtained by calculating the value of the test statistic and using Appendix Table 4 is due to differences in rounding in the hand calculations.

Example 10.14 Shopping Online at Work

The article **"Americans Spend Nearly Two Hours a Day Shopping Online at Work, Study Suggests" (cnbc.com, December 11, 2018)** summarizes data from a survey conducted by Finder.com. The title of this article suggests that 2 hours a day spent shopping online at work might be a plausible value for the mean time for the population of all American workers. The actual survey data can be found in the blog post **"Number of People Who Shop at Work" (finder.com, August 3, 2021)**. In a representative sample of 2020 adult Americans, the mean time spent online shopping while at work was 1.7 hours. The sample standard deviation (estimated from a relative frequency distribution that appeared in the blog post) was 0.8 hours. Do the survey data provide convincing evidence that the population mean time is different from the value of 2 hours suggested by the article headline?

A hypothesis test can be carried out to answer this question.

Understand the context ➤

1. μ = population mean time spent shopping online while at work
2. H_0: $\mu = 2$
3. H_a: $\mu \neq 2$
4. $\alpha = 0.05$

Formulate a plan ➤

5. $$t = \frac{\bar{x} - \text{hypothesized value}}{\frac{s}{\sqrt{n}}} = \frac{\bar{x} - 2}{\frac{s}{\sqrt{n}}}$$

6. This test requires a random sample and either a large sample size or a normal population distribution. The sample was selected to be representative of the population, so it is reasonable to regard the sample as if it were a random sample.

The following relative frequency distribution was given in the blog post.

Time Spent Shopping at Work	Relative Frequency
0 to 30 minutes	0.46
30 minutes to 1 hour	0.19
1 hour to 2 hours	0.17

(continued)

Copyright 2025 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s). Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

Time Spent Shopping at Work	Relative Frequency
2 hours to 5 hours	0.11
5 hours to 10 hours	0.04
10 hours to 20 hours	0.02
20+ hours	0.01

What Do You Think?

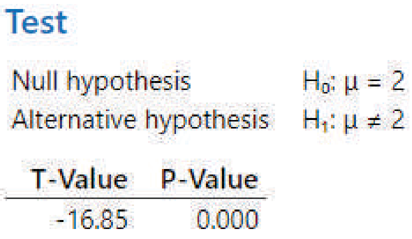
Based on the given relative frequency distribution, why would it not be reasonable to assume that the population distribution of times is approximately normal?

Because the given relative frequency distribution clearly shows that the distribution of times is not symmetric and is positively skewed, it is not reasonable to think that that population distribution is approximately normal. However, the sample size was large ($n = 2020$), so it is reasonable to proceed with the one-sample t test.

Do the work ➤

7. $$t = \frac{1.7 - 2}{\frac{0.8}{\sqrt{2020}}} = -16.85$$

8. Using technology to obtain the P-value for this two-sided test results in P-value $= 0.000$.



Test

Null hypothesis H₀: μ = 2
Alternative hypothesis H₁: μ ≠ 2

```
T-Value  P-Value
-16.85   0.000
```

Interpret the results ➤

9. Because the P-value is less than α, the null hypothesis is rejected. There is convincing evidence that the mean time spent shopping at work is not 2 hours. Because 2 hours is not a plausible value for the population mean, the article title "Americans Spend Nearly Two Hours a Day Shopping Online at Work, Study Suggests" may be misleading.

Statistical Versus Practical Significance

Carrying out a hypothesis test amounts to deciding whether the value obtained for the test statistic would have been likely to occur when H_0 is true. When the value of the test statistic leads to rejection of H_0, you can say that the result is **statistically significant** at the chosen significance level α. The finding of statistical significance means that the observed deviation from what was expected when H_0 is true cannot reasonably be explained by only chance variation. However, statistical significance is not the same as concluding that the true situation differs from what the null hypothesis states in any practical sense. That is, even after H_0 has been rejected, the data may suggest that there is no *practical* difference between the actual value of the population characteristic and what the null hypothesis states that value to be. This is illustrated in Example 10.15.

Example 10.15 "Significant" but Unimpressive Test Score Improvement

Understand the context ➤

Let μ denote the average score on a standardized test for all children in a large school district. The average score for all children in the United States is 100. District administrators are interested in testing H_0: $\mu = 100$ versus H_a: $\mu > 100$ using a significance level of 0.001.

Copyright 2025 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s). Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

A sample of 2500 children resulted in the values $n = 2500$, $\bar{x} = 101.0$, and $s = 15.0$. Then

$$t = \frac{101.10 - 100}{\frac{15}{\sqrt{2500}}} = 3.67$$

This is an upper-tailed test, so (using the z column of Appendix Table 4 because df $= 2499$) P-value $=$ area to the right of $3.67 \approx 0.000$. Because P-value < 0.001, you reject H_0. There is convincing evidence that the mean score for this school district is greater than 100.

Interpret the results ➤ However, with $n = 2500$ and a sample standard deviation of 15, the point estimate $\bar{x} = 101.0$ is likely to be very close to the actual value of μ. This means that it looks as though H_0 was rejected because $\mu \approx 101$ rather than 100. From a practical point of view, a 1-point difference may not be important. A statistically significant result does not necessarily mean that there are any practical consequences.

Exercises 10.43–10.62

● Data set available online

10.43 Give as much information as you can about the P-value of a t test in each of the following situations:

a. Upper-tailed test, df $= 8$, $t = 2.0$
b. Upper-tailed test, $n = 14$, $t = 3.2$
c. Lower-tailed test, df $= 10$, $t = -2.4$
d. Lower-tailed test, $n = 22$, $t = -4.2$

10.44 Give as much information as you can about the P-value of a t test in each of the following situations:

a. Two-tailed test, df $= 15$, $t = -1.6$
b. Two-tailed test, $n = 16$, $t = 1.6$
c. Two-tailed test, $n = 16$, $t = 6.3$

10.45 Give as much information as you can about the P-value of a t test in each of the following situations:

a. Two-tailed test, df $= 9$, $t = 0.73$
b. Upper-tailed test, df $= 10$, $t = -0.5$
c. Lower-tailed test, $n = 20$, $t = -2.1$
d. Lower-tailed test, $n = 20$, $t = -5.1$
e. Two-tailed test, $n = 40$, $t = 1.7$

10.46 Paint used to paint lines on roads must reflect enough light to be clearly visible at night. Suppose μ denotes the mean reflectometer reading (in lumens per square meter) for a new type of paint. It is recommended that the mean reading be greater than 100. A test of H_0: $\mu = 100$ versus H_a: $\mu > 100$ based on a sample of 15 observations resulted in $t = 3.2$. What conclusion is appropriate at each of the following significance levels?

a. $\alpha = 0.05$
b. $\alpha = 0.01$
c. $\alpha = 0.001$

10.47 A ball point pen has been designed so that actual mean writing lifetime under controlled conditions (involving the use of a writing machine) is at least 10 hours. A random sample of 18 pens is selected, the writing lifetime of each is determined, and a normal probability plot of the resulting data supports the use of a one-sample t test. The hypotheses of interest are H_0: $\mu = 10$ versus H_a: $\mu < 10$.

a. If $t = -2.3$ and $\alpha = 0.05$, what conclusion is appropriate?
b. If $t = -1.83$ and $\alpha = 0.01$, what conclusion is appropriate?
c. If $t = 0.47$, what conclusion is appropriate?

10.48 The United State Golf Association specifies that golf balls used in tournament play must have a minimum diameter of 1.68 inches. What conclusion is appropriate when testing H_0: $\mu = 1.68$ versus H_a: $\mu \neq 1.68$ inch each of the following situations:

a. $n = 13$, $t = 1.6$, $\alpha = 0.05$
b. $n = 13$, $t = -1.6$, $\alpha = 0.05$
c. $n = 25$, $t = -2.6$, $\alpha = 0.01$
d. $n = 25$, $t = -3.6$

10.49 The paper **"Playing Active Video Games Increases Energy Expenditure in Children" (*Pediatrics* [2009]: 534–539)** describes an interesting investigation of the possible cardiovascular benefits of active video games. Mean heart rate for healthy children age 10 to 13 after walking on a treadmill at 2.6 km/hour for 6 minutes is 98 beats per minute (bpm). For each of 14 children, heart rate was measured after 15 minutes of playing Wii Bowling. The resulting sample mean and standard deviation were 101 bpm and 15 bpm, respectively. For purposes of this exercise, assume that it is reasonable to regard the sample of children as representative of children age 10 to 13 and that the distribution of heart rates after 15 minutes of Wii Bowling is approximately normal.

a. Does the sample provide convincing evidence that the mean heart rate after 15 minutes of Wii Bowling is different from the known mean heart rate after 6 minutes walking on the treadmill? Carry out a hypothesis test using $\alpha = 0.01$.

b. The known resting mean heart rate for children in this age group is 66 bpm. Is there convincing evidence that the mean heart rate after Wii Bowling for 15 minutes is higher than the known mean resting heart rate for children of this age? Use $\alpha = 0.01$. (Hint: See Example 10.15.)

c. Based on the outcomes of the tests in Parts (a) and (b), write a paragraph comparing the benefits of treadmill walking and Wii Bowling in terms of raising heart rate over the resting heart rate.

10.50 The paper **"Privacy Policies for Apps Targeted Toward Youth: Descriptive Analysis of Readability" (*JIMR MHEALTH and UHEALTH* [2018]: 1–11)** describes a study of the readability of the privacy policies for apps that are targeted toward young people. The reading grade level (the grade level at which a person would be expected to be able to read and understand the text without help) was measured for the privacy policies of 34 apps targeted toward young people that were available for purchase. The sample mean reading grade level was reported to be 12.5 (meaning that someone would need to be reading at grade level 12.5 [senior in high school] to read and understand the policy). The sample standard deviation was 1.8. Suppose that it is reasonable to regard this sample of 34 as a random sample of apps available for purchase that are targeted toward young people. The average reading grade level for adults in the United States is thought to be 8.0. Do these data provide convincing evidence that the mean reading level of the privacy policies for apps targeted toward young people is greater than the mean reading level of adult Americans? Test the relevant hypotheses using a significance level of $\alpha = 0.01$.

10.51 The paper **"Lead in Spices, Herbal Remedies, and Ceremonial Powders Sampled in Home Investigation for Children with Elevated Blood Lead Levels—North Carolina 2011–2018" (*Morbidity and Mortality Weekly Report*, U.S. Department of Health and Human Services, November 23, 2018)** summarizes data from 61 young children living in a particular county in North Carolina who had their blood lead levels measured as part of a routine healthcare visit. The sample mean blood level was 17 μg/dL and the sample standard deviation was 7.0 μg/dL. Suppose that it is reasonable to regard this sample as representative of young children living in this county. Do these data provide convincing evidence that the mean blood lead level for young children in this county is greater than 10 μg/dL? Carry out a hypothesis test using $\alpha = 0.01$.

10.52 The paper referenced in the previous exercise also studied lead levels in spices found in homes in North Carolina. The spice turmeric was tested and in specimens from 34 homes, the sample mean lead level was reported to be 66.4 mg/kg and the sample standard deviation was 206.6 mg/kg. For purposes of this exercise, you can assume that this sample is representative of turmeric found in homes in North Carolina.

a. Explain why it is not reasonable to think that the distribution of lead level in turmeric is approximately normally distributed?

b. Even though it is not reasonable to assume a normal population distribution, is it still appropriate to use a one-sample *t* test to determine if there is convincing evidence that the mean lead concentration in turmeric found in homes in North Carolina is greater than 60 mg/kg? Explain why or why not.

c. If appropriate, carry out a hypothesis test to determine if there is convincing evidence that the mean lead concentration in turmeric found in homes in North Carolina is greater than 60 mg/kg. Use a significance level of 0.05.

10.53 The report **"NACE Salary Survey Executive Summary" (National Association of Colleges and Employers, Winter 2022)** states that the mean yearly salary offer for students graduating with mathematics and science degrees in 2021 was \$63,316. Suppose that a random sample of 50 math and science graduates at a large university who received job offers resulted in a mean offer of \$64,000 and a standard deviation of \$3,300. Do the sample data provide strong support for the claim that the mean salary offer for math and science graduates of this university is higher than the 2021 national average of \$63,316? Test the relevant hypotheses using $\alpha = 0.05$.

10.54 The report **"Majoring in Money: How American College Students Manage Their Finances" (Sallie Mae, 2016, news/salliemae.com)** includes data from a survey of college students. Each person in a representative sample of 793 college students was asked if they had one or more credit cards and if so, whether they paid their balance in full each month. There were 500 who paid in full each month. For this sample of 500 students, the sample mean credit card balance was reported to be \$825. The sample standard deviation of the credit card balances for these 500 students was not reported, but for purposes of this exercises, suppose that it was \$200. Is there convincing evidence that college students who pay their credit card balance in full each month have mean balance that is lower than \$906, the value reported for all college students with credit cards? Carry out a hypothesis test using a significance level of 0.01.

Copyright 2025 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s). Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

10.55 The authors of the paper **"Changes in Quantity, Spending, and Nutritional Characteristics of Adult, Adolescent and Child Urban Corner Store Purchases After an Environmental Intervention" (*Preventative Medicine* [2015]: 81–85)** wondered if increasing the availability of healthy food options would also increase the amount people spend at the corner store. They collected data from a representative sample of 5949 purchases at corner stores in Philadelphia after the stores increased their healthy food options. The sample mean amount spent for this sample of purchases was \$2.86 and the sample standard deviation was \$5.40.

a. Notice that for this sample, the sample standard deviation is greater than the sample mean. What does this tell you about the distribution of purchase amounts?

b. Before the stores increased the availability of healthy foods, the population mean total amount spent per purchase was thought to be about \$2.80. Do the data from this study provide convincing evidence that the population mean amount spent per purchase is greater after the change to increase healthy food options? Carry out a hypothesis test with a significance level of 0.05.

10.56 ● Medical research has shown that repeated wrist extension beyond 20 degrees increases the risk of wrist and hand injuries. Each of 24 students at Cornell University used a proposed new computer mouse design. While using the mouse, each student's wrist extension was recorded. Data consistent with summary values given in the paper **"Comparative Study of Two Computer Mouse Designs" (Cornell Human Factors Laboratory Technical Report RP7992)** are given. Use these data to test the hypothesis that the mean wrist extension for people using this new mouse design is greater than 20 degrees. Are any assumptions required in order for it to be appropriate to generalize the results of your test to the population of Cornell students? To the population of all university students? (Hint: See Example 10.13.)

27 28 24 26 27 25 25 24 24 24 25 28
22 25 24 28 27 26 31 25 28 27 27 25

10.57 A comprehensive study conducted by the National Institute of Child Health and Human Development tracked more than 1000 children from an early age through elementary school **(*The New York Times*, November 1, 2005)**. The study concluded that children who spent more than 30 hours a week in child care before entering school tended to score higher in math and reading when they were in the third grade. The researchers cautioned that the findings should not be a cause for alarm because the effects of child care were found to be small.

Explain how the difference between the sample mean math score for third graders who spent long hours in child care and the known overall mean for third graders could be small but the researchers could still reach the conclusion that the mean for the child care group is significantly higher than the overall mean for third graders. (Hint: See discussion of statistical versus practical significance.)

10.58 The paper **"Reducing the Use of Screen Electronic Devises in the Evening is Associated with Improved Sleep and Daytime Vigilance in Adolescents" (*Sleep Research Society* [2019]: 1–10)** describes a study of the negative effects of screen time before going to bed. The researchers measured sleep time for a sample of people age 12 to 19 years old at the beginning of the study and again after a period of time when participants refrained from using electronic devices after 9 p.m. They then recorded the increase in sleep time for each participant. The sample mean increase in sleep time was 17 minutes and the sample standard deviation of the increases in sleep time was 35.5. The increase in sleep times ranged from about −45 minutes (corresponding to a decrease of 45 minutes) to about 180 minutes.

a. Based on the given sample mean and standard deviation, is it reasonable to think that the distribution of sleep time increases is approximately normal? Explain why or why not.

b. Suppose that you wanted to use the sample data to decide if there is convincing evidence that the mean increase in sleep time for people in this age group who refrain from using electronic devices after 9 p.m. is greater than 15 minutes. What would you need to know to determine if the one-sample *t* test is an appropriate method?

10.59 Refer to the study description and sample statistics given in the previous exercise.

a. The sample size in this study was 315. Carry out a hypothesis test to decide if there is convincing evidence that the mean increase in sleep time for people in this age group who refrain from using electronic devices after 9 p.m. is greater than 15 minutes. Use a significance level of 0.05.

b. Now suppose that the sample standard deviation had been 20.5 rather than 35.5. Carry out a hypothesis test to decide if there is convincing evidence that the mean increase in sleep time for people in this age group who refrain from using electronic devices after 9 p.m. is greater than 15 minutes. Use a significance level of 0.05.

c. Explain why the null hypothesis was rejected in the test in Part (b) but not in the test of Part (a).

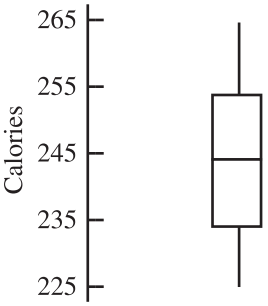
10.60 The paper titled **"Music for Pain Relief" (*The Cochrane Database of Systematic Reviews,* April 19, 2006)** concluded, based on a review of 51 studies of the effect of music on pain intensity, that

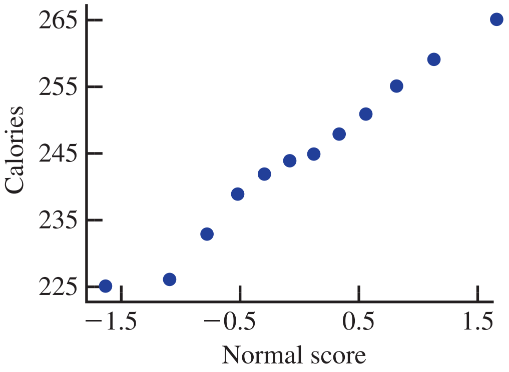
> "Listening to music reduces pain intensity levels … However, the magnitude of these positive effects is small, the clinical relevance of music for pain relief in clinical practice is unclear."

Are the authors of this paper claiming that the pain reduction attributable to listening to music is not statistically significant, not practically significant, or neither statistically nor practically significant? Explain.

10.61 ● Many consumers pay careful attention to stated nutritional contents on packaged foods when making purchases. It is therefore important that the information on packages be accurate. A random sample of $n = 12$ frozen dinners of a certain type was selected from production during a particular period, and the calorie content of each one was determined. Here are the resulting observations, along with a boxplot and normal probability plot:

255	244	239	242	265	245	259	248
225	226	251	233				





a. Is it reasonable to test hypotheses about mean calorie content μ by using a t test? Explain why or why not.

b. The stated calorie content is 240. Does the boxplot suggest that actual average content differs from the stated value? Explain your reasoning.

c. Carry out a formal test of the hypotheses suggested in Part (b).

10.62 ● Much concern has been expressed regarding the practice of using nitrates as meat preservatives. In one study involving possible effects of these chemicals, bacteria cultures were grown in a medium containing nitrates. The rate of uptake of radio-labeled amino acid (in dpm, disintegrations per minute) was then determined for each culture, yielding the following observations:

7251	6871	9632	6866	9094	5849	8957	7978
7064	7494	7883	8178	7523	8724	7468	

Suppose that it is known that the mean rate of uptake for cultures without nitrates is 8000.

Do the data suggest that the addition of nitrates results in a decrease in the mean rate of uptake? Test the appropriate hypotheses using a significance level of 0.10.

Section 10.5 Power and Probability of Type II Error

This chapter introduced test procedures for testing hypotheses about population characteristics, such as μ and p. What characterizes a "good" test procedure? It makes sense to think that a good test procedure is one that has both a small probability of rejecting the null hypothesis when it is true (a Type I error) and a high probability of rejecting the null hypothesis when it is false.

The test procedures presented in this chapter allow you to directly control the probability of rejecting a true null hypothesis by your choice of the significance level α. But what about the probability of rejecting the null hypothesis when it is false? Several factors influence this probability. Let's begin by considering an example.

Suppose that the student body president at a university is interested in studying the amount of money that students spend on textbooks each semester. The director of the financial aid office believes that the average amount spent on books is \$500 per semester and uses this figure to determine the amount of financial aid for which a student is eligible. The student

Copyright 2025 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s). Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

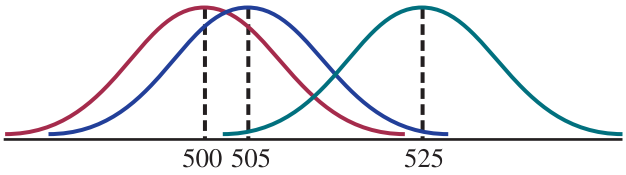
body president plans to ask each individual in a random sample of students how much they spent on books this semester. The resulting data will then be used to test

$$H_0\colon \mu = 500 \quad \text{versus} \quad H_a\colon \mu > 500$$

using a significance level of 0.05. If the actual mean is 500 (or less than 500), the correct decision is to fail to reject the null hypothesis. Incorrectly rejecting the null hypothesis is a Type I error. On the other hand, if the actual mean is 525 or 505 or even 501, the correct decision is to reject the null hypothesis. Not rejecting the null hypothesis is a Type II error. How likely is it that the null hypothesis will in fact be rejected?

If the actual mean is 501, the probability that you will reject $H_0\colon \mu = 500$ is not very great. This is because when you carry out the test, you consider the sample mean and ask if it looks like what you would expect to see if the population mean were 500. As illustrated in Figure 10.4, if the actual mean is greater than but very close to 500, chances are that the sample mean will look pretty much like what you would expect to see if the population mean were 500, and you will not be convinced that the null hypothesis should be rejected. If the true mean is 525, it is less likely that the sample will be mistaken for a sample from a population with mean 500 because sample means will tend to cluster around 525. In this case, it is more likely that you will correctly reject H_0. If the actual mean is 550, rejection of H_0 is even more likely.

Figure 10.4
Sampling distribution of $\bar{x}$ for $\mu = 500$, 505, and 525.



When you consider the probability of rejecting the null hypothesis, you are looking at what statisticians refer to as the **power** of the test.

Definition

Power of a test: The probability of rejecting the null hypothesis.

What Do You Think?

When the null hypothesis is false, why would you want the power of a test to be high?

From the previous discussion, it should be apparent that when a hypothesis about a population mean is being tested, the power of the test depends on the actual value of the population mean, μ. Because the actual value of μ is unknown (if you knew the value of μ you wouldn't be doing the hypothesis test!), you cannot know what the power is for the actual value of μ. However, it is possible to gain some insight into the power of a test by looking at a number of "what if" scenarios. For example, you might ask, What is the power if the actual mean is 525? or What is the power if the actual mean is 505? and so on. You can determine the power at $\mu = 525$, the power at $\mu = 505$, and the power at any other value of interest. Although it is technically possible to consider power when the null hypothesis is true, an investigator is usually concerned about the power only at values for which the null hypothesis is false.

In general, when testing a hypothesis about a population characteristic, there are three factors that influence the power of the test:

1. The size of the difference between the actual value of the population characteristic and the hypothesized value (the value that appears in the null hypothesis).
2. The choice of significance level, α.
3. The sample size.

Effect of Various Factors on the Power of a Test

1. The larger the size of the difference between the hypothesized value and the actual value of the population characteristic, the greater the power of the test.
2. The larger the significance level, α, the greater the power of the test.
3. The larger the sample size, the greater the power of the test.

Consider each of the statements in the box above. The first statement has already been discussed in the context of the textbook example. Because power is the probability of rejecting the null hypothesis, it makes sense that the power will be greater when the actual value of a population characteristic is quite different from the hypothesized value than when it is close to the hypothesized value.

The effect of significance level on power is not as obvious. To understand the relationship between power and significance level, it helps to consider the relationship between power and β, the probability of a Type II error.

When H_0 is false, power $= 1 - \beta$.

This relationship follows from the definitions of power and Type II error. A Type II error results from *not* rejecting a false H_0. Because power is the probability of rejecting H_0, it follows that *when H_0 is false*

$$\begin{aligned}\text{power} &= \text{probability of rejecting a false } H_0\\ &= 1 - \text{probability of not rejecting a false } H_0\\ &= 1 - \beta\end{aligned}$$

Recall from Section 10.2 that the choice of α, the Type I error probability, affects the value of β, the Type II error probability. Choosing a larger value for α results in a smaller value for β (and therefore a larger value for $1 - \beta$). In terms of power, this means that choosing a larger value for α results in a larger value for the power of the test. The larger the Type I error probability you are willing to tolerate, the more likely it is that the test will be able to detect any particular departure from H_0.

The third factor that affects the power of a test is the sample size. When H_0 is false, the power of a test is the probability that you will "detect" that H_0 is false and, based on the observed sample, reject H_0. Intuition suggests that you will be more likely to detect a departure from H_0 with a large sample than with a small sample. This is in fact the case—the larger the sample size, the higher the power.

For example, consider testing the hypotheses:

$$H_0\text{: } \mu = 500 \quad \text{versus} \quad H_a\text{: } \mu > 500$$

The observations about power imply the following:

1. If μ is greater than 500, the power of a test based on a sample of size 100 is greater than the power of a test based on a sample of size 75 (assuming the same significance level).
2. If μ is greater than 500, the power of a test using a significance level of 0.05 is greater than the power of a test using a significance level of 0.01 (assuming the same sample size).
3. If μ is greater than 500, the power of the test is greater if the actual mean is 550 than if the actual mean is 525 (assuming the same sample size and significance level).

As was mentioned previously in this section, it is impossible to calculate the *exact* power of a test because in practice you do not know the values of population characteristics. However, you can evaluate the power at a selected alternative value which would tell you whether the power would be high or low if this alternative value is the actual value.

The following optional subsection shows how Type II error probabilities and power can be evaluated for selected tests.

Calculating Power and Type II Error Probabilities for Selected Tests (Optional)

The test procedures presented in this chapter are designed to control the probability of a Type I error (rejecting H_0 when H_0 is true) at the desired significance level α. However, little has been said so far about calculating the value of β, the probability of a Type II error (not rejecting H_0 when H_0 is false). In this section, the calculation of β and power for the hypothesis tests introduced in this chapter is considered.

When you carry out a hypothesis test, you specify the desired value of α, the probability of a Type I error. The probability of a Type II error, β, is the probability of not rejecting H_0 even though it is false. Suppose that you are testing

$$H_0\colon\ \mu = 1.5 \quad \text{versus} \quad H_a\colon\ \mu > 1.5$$

Because you do not know the actual value of μ, you cannot calculate the actual value of β. However, the probability of Type II error can be investigated by calculating β for several different potential values of μ, such as $\mu = 1.55$, $\mu = 1.6$, and $\mu = 1.7$. Once a value of β has been determined, the power of the test for the corresponding alternative value is just $1 - \beta$.

Example 10.16 Calculating Power

An airline claims that the mean time on hold for callers to its customer service phone line is 1.5 minutes. You might investigate this claim by testing

$$H_0\colon\ \mu = 1.5 \quad \text{versus} \quad H_a\colon\ \mu > 1.5$$

where μ is the actual mean customer hold time. A random sample of $n = 36$ calls is to be selected, and the resulting data will be used to reach a conclusion.

Suppose that the standard deviation of hold time (σ) is known to be 0.20 minutes and that a significance level of 0.01 is to be used. The test statistic for this test would be

$$z = \frac{\bar{x} - 1.5}{\dfrac{0.20}{\sqrt{n}}} = \frac{\bar{x} - 1.5}{\dfrac{0.20}{\sqrt{36}}} = \frac{\bar{x} - 1.5}{0.0333}$$

The inequality in H_a implies that

P-value = area under z curve to the right of calculated z

Using technology or Appendix Table 2, it is easily verified that the z critical value 2.33 captures an upper-tail z curve area of 0.01. This means that P-value ≤ 0.01 only when $z \geq 2.33$. This is equivalent to the decision rule

reject H_0 if calculated $z \geq 2.33$

which becomes

$$\text{reject } H_0 \text{ if } \frac{\bar{x} - 1.5}{0.0333} \geq 2.33$$

Solving this inequality for $\bar{x}$ results in

$$\bar{x} \geq 1.5 + 2.33(0.0333)$$

or

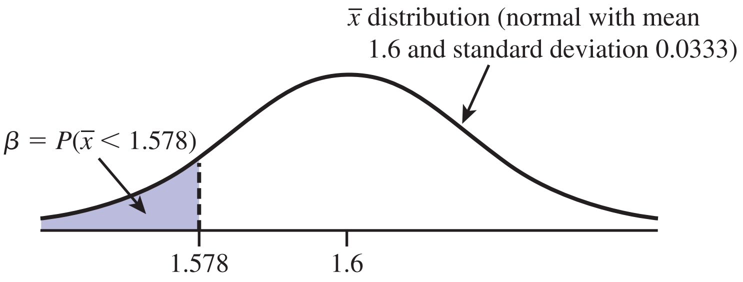
$$\bar{x} \geq 1.578$$

So if $\bar{x} \geq 1.578$, you will reject H_0, and if $\bar{x} < 1.578$, you will fail to reject H_0. This decision rule corresponds to $\alpha = 0.01$.

Suppose now that $\mu = 1.6$ (so that H_0 is false). A Type II error will then occur if $\bar{x} < 1.578$. What is the probability that this occurs? If $\mu = 1.6$, the sampling distribution of $\bar{x}$ is approximately normal, centered at 1.6, and has a standard deviation of 0.0333. The probability of observing an $\bar{x}$ value less than 1.578 can then be determined by finding an area under a normal curve with mean 1.6 and standard deviation 0.0333, as illustrated in Figure 10.5.

Copyright 2025 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s). Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

Figure 10.5
β for $\mu = 1.6$ in Example 10.16.



Because the curve in Figure 10.5 is not the standard normal (z) curve, you must first convert to a z score before using Appendix Table 2 to find the area, or you could use technology to find the area. Here,

$$z \text{ score for } 1.578 = \frac{1.578 - \mu_{\bar{x}}}{\sigma_{\bar{x}}} = \frac{1.578 - 1.6}{0.0333} = -0.66$$

and

area under z curve to left of $-0.66 = 0.2546$

So, if $\mu = 1.6$, $\beta = 0.2546$. This means that if μ is 1.6, about 25% of all samples would still result in $\bar{x}$ values less than 1.578, and you would not reject H_0.

The power of the test at $\mu = 1.6$ is then

$$\begin{aligned}(\text{power at } \mu = 1.6) &= 1 - (\beta \text{ when } \mu \text{ is } 1.6)\\ &= 1 - 0.2546\\ &= 0.7454\end{aligned}$$

This means that if the actual mean is 1.6, the probability of rejecting H_0: $\mu = 1.5$ in favor of H_a: $\mu > 1.5$ is 0.7454. If μ is 1.6 and the test is used repeatedly with random samples selected from the population, in the long run about 75% of the samples will result in the correct conclusion to reject H_0.

Now consider β and power when $\mu = 1.65$. The normal curve in Figure 10.5 would then be centered at 1.65. Because β is the area to the left of 1.578 and the curve has shifted to the right, β decreases. Converting 1.578 to a z score and using Appendix Table 2 or using technology gives $\beta = 0.0154$. Also,

$$(\text{power at } \mu = 1.65) = 1 - 0.0154 = 0.9846$$

As expected, the power at $\mu = 1.65$ is greater than the power at $\mu = 1.6$ because 1.65 is farther from the hypothesized value of 1.5.

Statistical software packages and some graphing calculators can calculate the power for specified values of σ, α, n, and the difference between the actual and hypothesized values of μ. The following Minitab output shows power calculations corresponding to those in Example 10.16:

```
1-Sample Z Test
Testing mean = null (versus > null)
Alpha = 0.01     Sigma = 0.2       Sample Size = 36
Difference           Power
0.10                0.7497
0.15                0.9851
```

The slight differences between the power values computed by Minitab and those previously obtained are due to rounding in Example 10.16.

The probability of a Type II error and the power for z tests concerning a population proportion are calculated in a similar manner.

Copyright 2025 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s). Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

Example 10.17 Power for Testing Hypotheses About Proportions

A package delivery service advertises that at least 90% of all packages brought to its office by 9 A.M. for delivery in the same city are delivered by noon that day. Let p denote the actual proportion of all such packages that are delivered by noon. The hypotheses of interest are

$$H_0\colon\ p = 0.9 \quad \text{versus} \quad H_a\colon\ p < 0.9$$

where the alternative hypothesis states that the company's claim is not true.

The value $p = 0.8$ represents a substantial departure from the company's claim. If the hypotheses are tested using significance level 0.01 and a sample of $n = 225$ packages, what is the probability that the departure from H_0 represented by this alternative value will go undetected?

At significance level 0.01, H_0 is rejected if P-value ≤ 0.01. For the case of a lower-tailed test, this is the same as rejecting H_0 if

$$z = \frac{\hat{p} - \mu_{\hat{p}}}{\sigma_{\hat{p}}} = \frac{\hat{p} - 0.9}{\sqrt{\dfrac{(0.9)(1-0.9)}{225}}} = \frac{\hat{p} - 0.9}{0.02} \leq -2.33$$

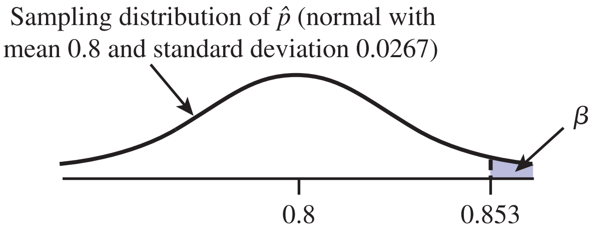
(Because -2.33 captures a lower-tail z curve area of 0.01, the smallest 1% of all z values satisfy $z \leq -2.33$.) This inequality is equivalent to $\hat{p} \leq 0.853$, so H_0 is *not* rejected if $\hat{p} > 0.853$. When $p = 0.8$, $\hat{p}$ has approximately a normal distribution with

$$\mu_{\hat{p}} = 0.8$$

$$\sigma_{\hat{p}} = \sqrt{\frac{(0.8)(1-0.8)}{225}} = 0.0267$$

Then β is the probability of obtaining a sample proportion greater than 0.853, as illustrated in Figure 10.6.

Figure 10.6
β for $p = 0.8$ in Example 10.17.



Converting to a z score results in

$$z = \frac{0.853 - 0.8}{0.0267} = 1.99$$

and Appendix Table 2 or using technology results in

$$\beta = 1 - 0.9767 = 0.0233$$

When $p = 0.8$ and a significance level of 0.01 is used, less than 3% of all samples of size $n = 225$ will result in a Type II error. The power of the test for $p = 0.8$ is $1 - 0.0233 = 0.9767$. This means that the probability of rejecting $H_0\colon p = 0.9$ in favor of $H_a\colon p < 0.9$ when p is really 0.8 is 0.9767, which is quite high.

β and Power for the t Test (Optional)

The power and β values for t tests can be approximated by using a set of curves constructed for this purpose or by using appropriate software. As with the z test, the value of β depends not only on the actual value of μ but also on the selected significance level

Copyright 2025 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s). Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

α. β increases as α is made smaller. In addition, β depends on the number of degrees of freedom, $n - 1$. For any fixed significance level α, it should be easier for the test to detect a specific departure from H_0 when n is large than when n is small. For a fixed alternative value, β decreases as $n - 1$ increases.

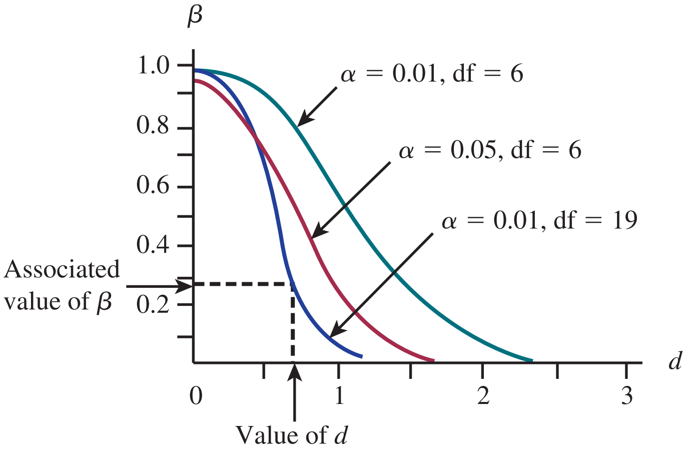
Unfortunately, there is one other quantity on which β depends: the population standard deviation σ. As σ increases, so does $\sigma_{\bar{x}}$. This in turn makes it more likely that an $\bar{x}$ value far from μ will be observed just by chance, resulting in an incorrect conclusion. Once α is specified and n is fixed, the determination of β at a particular alternative value of μ requires that a value of σ be chosen, because each different value of σ yields a different value of β. (This did not present a problem with the z test because when using a z test, the value of σ is known.) If it is possible to specify a range of plausible values for σ, then using the largest such value will give a conservative estimate of β (one on the high side) and a conservative estimate of power (one on the low side).

Figure 10.7 shows three different β curves for a one-tailed t test (appropriate for H_a: $\mu >$ hypothesized value or for H_a: $\mu <$ hypothesized value). A more complete set of curves for both one-tailed and two-tailed tests when $\alpha = 0.05$ and when $\alpha = 0.01$ appears in Appendix Table 5. To determine β, first calculate the quantity

$$d = \frac{|\text{alternative value} - \text{hypothesized value}|}{\sigma}$$

Then locate d on the horizontal axis, move directly up to the curve for $n - 1$ df, and move over to the vertical axis to obtain an estimate of β.

Figure 10.7
β curves for the one-tailed t test.



Example 10.18 β and Power for t Tests

Consider testing

$$H_0\text{: } \mu = 100 \quad \text{versus} \quad H_a\text{: } \mu > 100$$

and focus on the alternative value $\mu = 110$. Suppose that $\sigma = 10$, the sample size is $n = 7$, and a significance level of 0.01 has been selected. For $\sigma = 10$,

$$d = \frac{|110 - 100|}{10} = \frac{10}{10} = 1$$

Figure 10.7 (using df $= 7 - 1 = 6$) gives $\beta \approx 0.6$. This means that if $\sigma = 10$, the significance level is 0.01 and $n = 7$, when $\mu = 110$, roughly 60% of all samples result in an incorrect decision to not reject H_0! Equivalently, the power of the test for $\mu = 110$ is only $1 - 0.6 = 0.4$. The probability of rejecting H_0 when $\mu = 110$ is not very large.

If a 0.05 significance level is used instead, then $\beta \approx 0.3$, which is still rather large. Using a 0.01 significance level with $n = 20$ (df $= 19$) yields, from Figure 10.7, $\beta \approx 0.05$. At the alternative value $\mu = 110$, for $\sigma = 10$, a significance level of 0.01, and $n = 20$, β is less than for a significance level of 0.05 with $n = 7$. Substantially increasing n counterbalances using the smaller α.

Copyright 2025 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s). Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

Now consider the alternative $\mu = 105$, again with $\sigma = 10$, so that

$$d = \frac{|105 - 100|}{10} = \frac{5}{10} = 0.5$$

Then, from Figure 10.7, $\beta \approx 0.95$ when $\alpha = 0.01$, $n = 7$; $\beta \approx 0.7$ when $\alpha = 0.05$, $n = 7$; and $\beta \approx 0.65$ when $\alpha = 0.01$, $n = 20$. These values of β are all quite large. With $\sigma = 10$, $\mu = 105$ is too close to the hypothesized value of 100 for any of these three tests to have a good chance of detecting this departure from H_0. A substantial decrease in β would require using a much larger sample size. For example, from Appendix Table 5, $\beta \approx 0.08$ when $\alpha = 0.05$ and $n = 40$.

The curves in Figure 10.7 also give β when testing H_0: $\mu = 100$ versus H_a: $\mu < 100$. If the alternative value $\mu = 90$ is of interest and $\sigma = 10$,

$$d = \frac{|90 - 100|}{10} = \frac{10}{10} = 1$$

and values of β are the same as those given in the first paragraph of this example.

Because curves for only selected degrees of freedom appear in Appendix Table 5, other degrees of freedom require a visual approximation. For example, the 27-df curve (for $n = 28$) lies between the 19-df and 29-df curves, which do appear, and it is closer to the 29-df curve. This type of approximation is adequate because it is the general magnitude of β—large, small, or moderate—that is of primary concern.

Minitab can also evaluate power for the *t* test. For example, the following output shows Minitab calculations for power at $\mu = 110$ for samples of size 7 and 20 when $\alpha = 0.01$. The corresponding approximate values from Appendix Table 5 found in Example 10.18 are fairly close to the Minitab values.

```
1-Sample t Test
Testing mean = null (versus > null)
Calculating power for mean = null + 10
Alpha = 0.01     Sigma = 10
  Sample Size    Power
            7    0.3968
           20    0.9653
```

The β curves in Appendix Table 5 are those for *t* tests. When the alternative value in H_a corresponds to a value of *d* relatively close to 0, β for a *t* test may be rather large. One might wonder whether there is another type of test that has the same level of significance α as does the *t* test and smaller values of β. The following result provides the answer to this question.

> When the population distribution is normal, the *t* test for testing hypotheses about μ has smaller β than does any other test procedure that has the same level of significance α.

Stated another way, among all tests with level of significance α, the *t* test makes β as small as it can possibly be when the population distribution is normal. In this sense, the *t* test is a best test. Statisticians have also shown that when the population distribution is not too far from a normal distribution, no test procedure can improve on the *t* test by very much (i.e., no test procedure can have the same α and substantially smaller β). However, when the population distribution is believed to be strongly nonnormal (heavy-tailed, highly skewed, or multimodal), the *t* test should not be used. One alternative method that could be used in this situation is introduced in Section 10.8.

Exercises 10.63–10.71

10.63 The power of a test is influenced by the sample size and the choice of significance level.

a. Explain how increasing the sample size affects the power (when significance level is held fixed).

b. Explain how increasing the significance level affects the power (when sample size is held fixed).

10.64 Water specimens are taken from water used for cooling as it is being discharged from a power plant into a river. It has been determined that as long as the mean temperature of the discharged water is at most 150°F, there will be no negative effects on the river's ecosystem. To investigate whether the plant is in compliance with regulations that prohibit a mean discharge water temperature above 150°F, a scientist will take 50 water specimens at randomly selected times and will record the water temperature of each specimen. The scientist will then use a z statistic

$$z = \frac{\bar{x} - 150}{\frac{\sigma}{\sqrt{n}}}$$

to decide between the hypotheses H_0: $\mu = 150$ and H_a: $\mu > 150$, where μ is the mean temperature of discharged water. Assume that σ is known to be 10.

a. Explain why use of the z statistic is appropriate in this setting.

b. Describe Type I and Type II errors in this context.

c. The rejection of H_0 when $z \geq 1.8$ corresponds to what value of α? (That is, what is the area under the z curve to the right of 1.8?)

d. Suppose that the actual value for μ is 153 and that H_0 is to be rejected if $z \geq 1.8$. Draw a sketch (similar to that of Figure 10.5) of the sampling distribution of $\bar{x}$, and shade the region that would represent β, the probability of making a Type II error.

10.65 Consider the water temperature situation described in the previous exercise.

a. For the hypotheses and test procedure described in the previous exercise, what is the value of β when $\mu = 153$. (Hint: See Example 10.16.)

b. For the hypotheses and test procedure described in the previous exercise, what is the value of β if $\mu = 160$?

c. What would be the conclusion of the test if H_0 is rejected when $z \geq 1.8$ and $\bar{x} = 152.4$? What type of error might have been made in reaching this conclusion?

10.66 Let μ denote the mean lifetime (in hours) for a certain type of battery under controlled laboratory conditions. A test of H_0: $\mu = 10$ versus H_a: $\mu < 10$ will be based on a sample of size 36. Suppose that σ is known to be 0.6, so $\sigma_{\bar{x}} = 0.1$. The appropriate test statistic is then

$$z = \frac{\bar{x} - 10}{0.1}$$

a. What is α for the test procedure that rejects H_0 if $z \leq -1.28$?

b. If the test procedure of Part (a) is used, find the value of β when $\mu = 9.8$, and interpret this error probability.

c. Without doing any calculation, explain how β when $\mu = 9.5$ compares to β when $\mu = 9.8$. Then check your assertion by finding the value β when $\mu = 9.5$.

d. What is the power of the test when $\mu = 9.8$? when $\mu = 9.5$?

10.67 The city council in a large city has become concerned about the trend toward exclusion of renters with children in apartments within the city. The housing coordinator has decided to select a random sample of 125 apartments and determine for each whether children are permitted. Let p be the proportion of all apartments that exclude children. If the city council is convinced that p is greater than 0.75, it will consider appropriate legislation.

a. If 102 of the 125 sampled apartments exclude renters with children, would a level 0.05 test lead you to the conclusion that more than 75% of all apartments exclude children?

b. What is the power of the test when $p = 0.8$ and $\alpha = 0.05$? (Hint: See Example 10.17.)

10.68 The amount of shaft wear after a fixed mileage was determined for each of seven randomly selected internal combustion engines, resulting in a mean of 0.0372 inch and a standard deviation of 0.0125 inch.

a. Assuming that the distribution of shaft wear is normal, use $\alpha = 0.05$ to test the hypotheses H_0: $\mu = 0.035$ versus H_a: $\mu > 0.035$.

b. Using $\sigma = 0.0125$, $\alpha = 0.05$, and Appendix Table 5, what is the approximate value of β, the probability of a Type II error, when $\mu = 0.04$? (Hint: See Example 10.18.)

c. What is the approximate power of the test when $\mu = 0.04$ and $\alpha = 0.05$?

10.69 Optical fibers are used in telecommunications to transmit light. Suppose current technology allows production of fibers that transmit light about 50 km. Researchers are trying to develop a new type of glass fiber that will increase this distance. In evaluating a new fiber, it is of interest to test H_0: $\mu = 50$

versus H_a: $\mu > 50$, with μ denoting the mean transmission distance for the new optical fiber.

a. Assuming $\sigma = 10$ and $n = 10$, use Appendix Table 5 to approximate β, the probability of a Type II error, for each of the given alternative values of μ for a test with significance level 0.05:

i. 52 ii. 55 iii. 60 iv. 70

b. What happens to β in each of the cases in Part (a) if σ is actually larger than 10? Explain your reasoning.

10.70 Let μ denote the mean diameter for ball bearings of a certain type. A test of H_0: $\mu = 0.5$ versus H_a: $\mu \neq 0.5$ will be based on a sample of n ball bearings. The diameter distribution is believed to be normal. Determine the approximate value of β in each of the following cases:

a. $n = 15, \alpha = 0.05, \sigma = 0.02, \mu = 0.52$
b. $n = 15, \alpha = 0.05, \sigma = 0.02, \mu = 0.48$
c. $n = 15, \alpha = 0.01, \sigma = 0.02, \mu = 0.52$

10.71 Use the information given in the previous exercise to determine the approximate value of β in each of the following cases.

a. $n = 15, \alpha = 0.05, \sigma = 0.02, \mu = 0.54$
b. $n = 15, \alpha = 0.05, \sigma = 0.04, \mu = 0.54$
c. $n = 20, \alpha = 0.05, \sigma = 0.04, \mu = 0.54$

Section 10.6 Interpreting and Communicating the Results of Statistical Analyses

The nine-step procedure for testing hypotheses used in this chapter provides a systematic approach for carrying out a complete test. However, you rarely see the results of a hypothesis test reported in publications in such a complete way.

Communicating the Results of Statistical Analyses

When summarizing the results of a hypothesis test, it is important that you include several things in the summary in order to provide all the relevant information. These are:

1. **Hypotheses.** Whether specified in symbols or described in words, it is important that both the null and the alternative hypotheses be clearly stated. If you are using symbols to define the hypotheses, be sure to describe them in the context of the problem at hand (for example, μ = population mean calorie intake).
2. **Test procedure.** You should be clear about what test procedure was used (for example, large-sample z test for proportions) and why you think it was reasonable to use this procedure. The plausibility of any required assumptions should be satisfactorily addressed.
3. **Test statistic and *p*-value.** Be sure to include the value of the test statistic and the associated *P*-value. Including the *P*-value allows a reader who may have chosen a different significance level to see whether they would have reached the same or a different conclusion.
4. **Conclusion in context.** Never end the report of a hypothesis test with the statement "I rejected (or did not reject) H_0." Always provide a conclusion that is in the context of the problem and that answers the original research question that the hypothesis test was designed to answer. Be sure also to indicate the level of significance used as a basis for the decision.

Interpreting the Results of Statistical Analyses

When the results of a hypothesis test are reported in a journal article or other published source, it is common to find only the value of the test statistic and the associated *P*-value accompanying the discussion of conclusions drawn from the data. Often, especially in newspaper articles, only sample summary statistics are given, with the conclusion immediately following. You may have to fill in some of the intermediate steps for yourself to see whether or not the conclusion is justified.

For example, the paper **"Serum Vitamin D Levels in Different Morphologic Forms of Age-Related Cataract" (*Acta Endrocronoligica* [2020]: 178–182)** summarizes the results of a study to determine if patients with age-related cataracts have a vitamin D deficiency. They recorded vitamin D levels for each person in two different samples. A vitamin D level less than 20 ng/mL is considered deficient. The two samples represented patients with two different types of age-related cataracts. The following is from the summary of results that appeared in the paper:

> Patients in Group 1 had mean vitamin D levels of 17.31 ± 13.30 ng/mL. Vitamin D levels in Group 2 patients were 13.34 ± 7.87 ng/mL. When the level of vitamin D insufficiency was accepted as 20 ng/mL, the mean serum vitamin D levels of the patients were not significantly different in Group 1 ($P = 0.31$; one-sample *t*-test) and statistically significantly lower in Group 2 patients ($P = 0.00$; one-sample *t*-test).

This suggests that the researchers considered the hypotheses H_0: $\mu = 20$ versus H_0: $\mu \neq 20$ for patients with the first type of cataracts, and did not reject the null hypothesis. For patients with the second type of cataracts, they considered the hypotheses H_0: $\mu = 20$ versus H_0: $\mu < 20$, and rejected the null hypothesis. The sample sizes were 26 for Group 1 and 53 for Group 2. Assuming that the two samples were representative of patients with these two types of cataracts and that the data distribution for the Group 1 sample was approximately normal, the one-sample t test was a reasonable choice for these hypothesis tests.

What to Look for in Published Data

Here are some questions to consider when you are reading a report that contains the results of a hypothesis test:

- What hypotheses are being tested? Are the hypotheses about a population mean, a population proportion, or some other population characteristic?
- Was the appropriate test used? Does the validity of the test depend on any assumptions about the sample or about the population from which the sample was selected? If so, are the assumptions reasonable?
- What is the *P*-value associated with the test? Was a significance level reported (as opposed to simply reporting the *P*-value)? Is the chosen significance level reasonable?
- Are the conclusions drawn consistent with the results of the hypothesis test?

For example, consider the following statement from the paper **"Didgeridoo Playing as Alternative Treatment for Obstructive Sleep Apnoea Syndrome" (*British Medical Journal* [2006]: 266–270)**: "We found that four months of training of the upper airways by didgeridoo playing reduces daytime sleepiness in people with snoring and obstructive apnoea syndrome." A didgeridoo is a wind instrument made from hollow wood that originated with the aboriginal tribes of Northern Australia. This statement was supported by data on a measure of daytime sleepiness called the Epworth scale. For the 14 participants in the study, the mean improvement in Epworth scale was 4.4 and the standard deviation was 3.7.

The paper does not indicate what test was performed or what the value of the test statistic was. It appears that the hypotheses of interest are H_0: $\mu = 0$ (no improvement) versus H_a: $\mu > 0$, where μ represents the mean improvement in Epworth score after 4 months of didgeridoo playing for all people with snoring and obstructive sleep apnoea. Because the sample size is not large, the one-sample t test would be appropriate if the sample can be considered a random sample and the distribution of Epworth scale improvement scores is approximately normal. If these assumptions are reasonable (something that was not addressed in the paper), the t test results in $t = 4.45$ and an associated P-value of 0.000. Because the reported P-value is so small H_0 would be rejected, supporting the conclusion in the paper that didgeridoo playing is an effective treatment.

A Word to the Wise: Cautions and Limitations

There are several things you should watch for when conducting a hypothesis test or when evaluating a written summary of a hypothesis test.

1. A hypothesis test can never show strong support for the null hypothesis. Make sure that you don't confuse "There is no reason to believe the null hypothesis is not true"

with the statement "There is convincing evidence that the null hypothesis is true." These are very different statements!

2. If you have complete information for the population, don't carry out a hypothesis test! It should be obvious that no test is needed to answer questions about a population if you have complete information and don't need to generalize from a sample, but people sometimes forget this fact. For example, in an article on growth in the number of prisoners by state, the ***San Luis Obispo Tribune*** **(August 13, 2001)** reported "California's numbers showed a statistically insignificant change, with 66 fewer prisoners at the end of 2000." The use of the term "statistically insignificant" implies some sort of statistical inference, which is not appropriate when a complete accounting of the entire prison population is known. Perhaps the author confused statistical and practical significance.
3. Don't confuse statistical significance with practical significance. When statistical significance has been established, be sure to step back and evaluate the result in light of its practical importance. For example, you may be convinced that the proportion who respond favorably to a proposed new medical treatment is greater than 0.4, the known proportion who responds favorably for the currently recommended treatments. But if the estimate of this proportion for the proposed new treatment is 0.405, is this of any practical interest? It might be if the proposed treatment is less costly or has fewer side effects, but in other cases it may not be of any real interest. Conclusions should always be interpreted in context.

Exercises 10.72–10.73

10.72 In 2006, Boston Scientific sought approval for a new heart stent (a medical device used to open clogged arteries) called the Liberte. This stent was being proposed as an alternative to a stent called the Express that was already on the market. The following excerpt is from an article that appeared in ***The Wall Street Journal*** **(August 14, 2008)**:

> Boston Scientific wasn't required to prove that the Liberte was 'superior' than a previous treatment, the agency decided—only that it wasn't 'inferior' to Express. Boston Scientific proposed—and the FDA okayed—a benchmark in which Liberte could be up to three percentage points worse than Express—meaning that if 6% of Express patients' arteries reclog, Boston Scientific would have to prove that Liberte's rate of reclogging was less than 9%. Anything more would be considered 'inferior.' . . . In the end, after nine months, the Atlas study found that 85 of the patients suffered reclogging. In comparison, historical data on 991 patients implanted with the Express stent show a 7% rate. Boston Scientific then had to answer this question: Could the study have gotten such results if the Liberte were truly inferior to Express?"

Assume a 7% reclogging rate for the Express stent. Explain why it would be appropriate for Boston Scientific to carry out a hypothesis test using the following hypotheses:

$$H_0: p = 0.10$$
$$H_a: p < 0.10$$

where p is the proportion of patients receiving Liberte stents that suffer reclogging. Be sure to address both the choice of the hypothesized value and the form of the alternative hypothesis in your explanation.

10.73 The article **"Eat Brownies for Better Bloodwork—And Other Easy Ways to Lower Triglycerides" (*Woman's World*, June 6, 2023)** included the following statements:

> Happy news: Indulging in gooey, fudgy brownies can keep your triglycerides low. The trick is simply replacing the butter or oil in your favorite recipe with pureed black beans. (The swap won't hamper the texture or taste!) According to an animal study done by the USDA, mice that ate a single serving of black beans a day lowered their triglycerides by 37%.

a. Based on these statements, do you think that the title of the article is appropriate? What questions would you want to ask the person who wrote the article?

b. The article includes a reference to a research study conducted by the U.S. Department of Agriculture. This study was published in the paper **"Cooked Black Turtle Beans Ameliorate Insulin Resistance and Restore Gut Microbiota in C57BL/6J Mice on High-Fat Diets"** that appeared in 2021 in the journal *Foods*. In this study, a sample of eight mice were fed a high-fat diet for 6 weeks and the decrease in triglycerides was measured. The "37% lower" figure from the article was based on comparing this group of eight mice to the value observed for mice who were on a high-fat diet but who did not consume

Copyright 2025 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s). Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

black beans, which was about 45 mg/dL (approximate value based on a graph that appeared in the paper). If the sample mean for the mice that ate a high-fat diet supplemented by a serving of black beans was 37% lower than 45, what is the approximate value of this sample mean?

c. An approximate value for the standard deviation of the triglyceride levels for mice in the sample who were fed a high-fat diet supplemented by black beans is 15 (approximated from a graph in the paper). Using μ to represent the mean triglyceride for mice that eat a high-fat diet supplemented with black beans, would it be reasonable to use a one-sample t test to test the hypotheses H_0: $\mu = 45$ versus H_a: $\mu < 45$? Justify your answer.

d. Researchers also looked at the mean triglyceride level of mice who were on a low-fat diet and concluded that this mean was lower than both the mean of the mice fed a high-fat diet with the black beans supplement and the mice fed a high-fat diet without the black bean supplement. Given this information and the information in Part (c), write a few sentences commenting on whether the recommendations in the *Woman's World* article are justified.

Section 10.7 Randomization Test and Exact Binomial Test for a Population Proportion (Optional)

In Section 10.3, you learned how data from a random sample could be used to carry out a large-sample z test for one proportion. For that test to be appropriate, the sample must be a random sample from a population (or is selected in a way that makes it reasonable to think that the sample is representative of a population). In addition, the distribution of the sample proportion, $\hat{p}$, should be approximately normal. It is reasonable to think that the distribution of $\hat{p}$ is approximately normal when the sample size is large ($np \geq 10$ and $n(1-p) \geq 10$), but the distribution of $\hat{p}$ is not always approximately normal when the sample size is small.

When the sample size is large, a z test makes use of the normal distribution to calculate a P-value. But even when the sampling distribution of $\hat{p}$ is well approximated by a normal distribution, the resulting P-value for the large-sample z test is still just an approximation to the actual P-value.

When the sample size is not large enough to assume that the sampling distribution of $\hat{p}$ is normal, there are other methods that can be used to approximate a P-value for the hypothesis test that do not require a large sample size. Two of these methods are a randomization test and an "exact" test that is based on the binomial distribution.

A Randomization Test for One Proportion

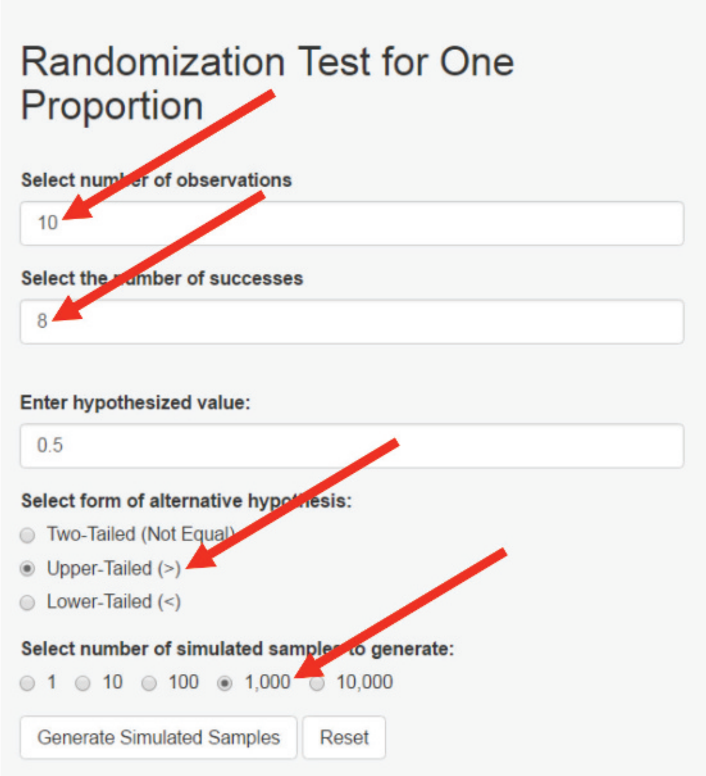
In a hypothesis test for one proportion, the null hypothesis is of the form H_0: $p =$ hypothesized value, where p is the population proportion and the hypothesized value is determined by the question of interest. For example, the null hypothesis might be H_0: $p = 0.50$.

Recall that in a hypothesis test, the P-value is a measure of how likely it would be to see something as extreme or more extreme as what was observed in the sample data if the null hypothesis were true. One way to approximate a P-value is to assume that the population proportion is in fact equal to the value specified in the null hypothesis and then simulate random samples from such a population. The distribution of these simulated sample proportions provides a sense of what values for the sample proportion would be expected when the null hypothesis is true and what values would be unlikely to occur. The distribution of simulated proportions also provides a way to find an approximate P-value. This distribution of simulated sample proportions is called a **randomization distribution**.

Example 10.19 Using a Shiny App to Create a Randomization Distribution

Suppose that when deciding if there is evidence that a majority of the students at a college are registered to vote, a random sample of 10 students includes 8 who are registered to vote. A majority of the students in the sample are registered, but does this mean it is reasonable to conclude that there is convincing evidence that a majority of *all* students are registered to vote? To answer this question, you could test the null hypothesis H_0: $p = 0.50$ against the alternative hypothesis H_a: $p > 0.50$.

Assuming that H_0: $p = 0.50$ is true, then values of $\hat{p}$ can be generated for many different simulated samples of size $n = 10$. Open the Shiny app called "Randomization Test for One Proportion," one of the Shiny web apps that accompany this text. These web apps are located at **www.cengage.com/Peck**. Enter 10 for the number of observations and 8 for the number of successes. The default hypothesized value for p is already 0.5. Choose "Upper-Tailed (>)" for the form of the alternative hypothesis, and request 1000 simulated samples. Click "Generate Simulated Samples."



Randomization Test for One Proportion

Select number of observations

10

Select the number of successes

8

Enter hypothesized value:

0.5

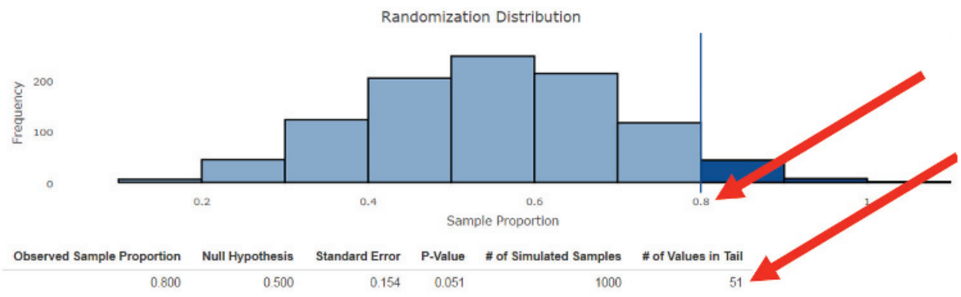
Select form of alternative hypothesis:

○ Two-Tailed (Not Equal)
● Upper-Tailed (>)
○ Lower-Tailed (<)

Select number of simulated samples to generate:

○ 1 ○ 10 ○ 100 ● 1,000 ○ 10,000

The Shiny app generates 1000 simulated random samples of size 10 from a population with $p = 0.50$ and provides a histogram of the values of the 1000 simulated sample proportions. The Shiny app also indicates where the observed value of the sample proportion, $\hat{p} = 0.8$, falls in the randomization distribution.



Observed Sample Proportion	Null Hypothesis	Standard Error	P-Value	# of Simulated Samples	# of Values in Tail
0.800	0.500	0.154	0.051	1000	51

The app reports that for this particular simulation, 51 of the 1000 simulated samples, or about 5.1%, were as or more extreme than what was observed in the actual sample. This value, $\frac{51}{1000} = 0.051$, is an approximation for the P-value for the one-sided test. The approximate P-value is also shown in the output produced by the app.

Based on this randomization distribution of 1000 simulated values of $\hat{p}$ when $n = 10$ and the null hypothesis H_0: $p = 0.5$ is true, the approximate P-value is $51/1000 = 0.051$. This P-value would then be compared to a selected significance level in order to make a decision about whether the null hypothesis should be rejected. For example, when using a significance level of 0.01, you would fail to reject H_0 because the P-value of 0.051 is greater than $\alpha = 0.01$. This means that even though 8 out of 10 students in the sample were registered to vote, it would not have been surprising to have seen this just by chance when the null hypothesis is true.

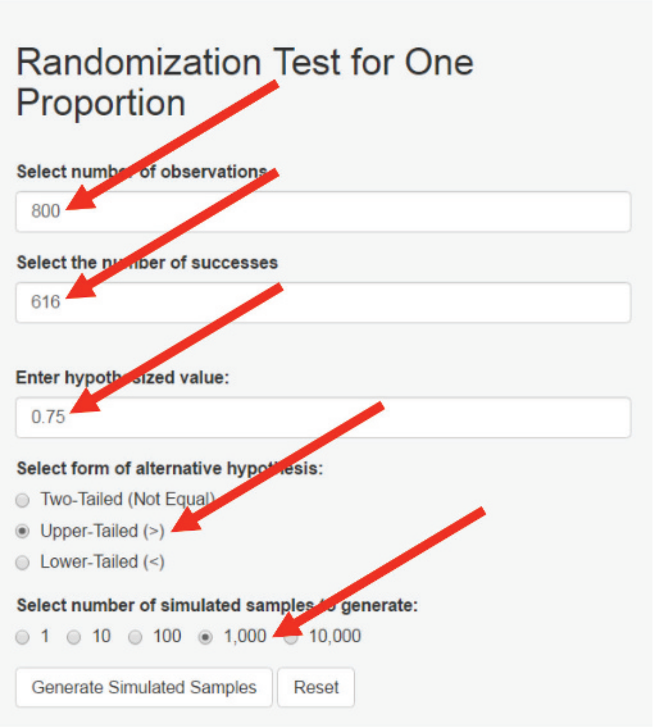
Example 10.20 Cell Phones Revisited

The paper referenced in Example 10.11 described a study of 2000 Canadians over the age of 18 that was carried out by Microsoft. Study participants were asked whether the following statement described them: "When nothing is occupying my attention, the first thing I do is reach for my phone." Of the study participants in the age group 18 to 24 years old, 77% responded "yes" to this question. In Example 10.11, it was assumed that the 77% was based on a representative sample of 800 Canadians age 18 to 24 years. A large-sample z test was used to decide if the sample data support the claim that more than 75% of all Canadians in this age group would respond "yes" when asked if the given statement describes them. In this example, the relevant hypotheses will be tested using a randomization test.

Let p represent the proportion of Canadians age 18 to 24 years who would respond "yes" for the population represented by the sample of 800. The appropriate hypotheses for the test are H_0: $p = 0.75$ and H_a: $p > 0.75$.

Open the Shiny app called "Randomization Test for One Proportion." This is one of the Shiny web apps that accompany this text. These web apps are located at **www.cengage.com/Peck**. Enter 800 for the Number of observations, and 616 (77% of 800) for the Number of successes. The default hypothesized value for p is 0.5, so change it to 0.75. Choose "Upper-Tailed (>)" for the form of the alternative hypothesis, and request 1000 simulated samples.

Confirm that the value of $\hat{p} = 616/800 = 0.770$ appears below "Observed Sample Proportion" in the simulation output. Notice that the value of "# of Values in Tail" in the simulation output shown in the accompanying figure is 107, and that the P-value is $107/1000 = 0.107$. This represents a randomization test approximation for the P-value for the hypothesis test.



Randomization Test for One Proportion

Select number of observations

800

Select the number of successes

616

Enter hypothesized value:

0.75

Select form of alternative hypothesis:

○ Two-Tailed (Not Equal)

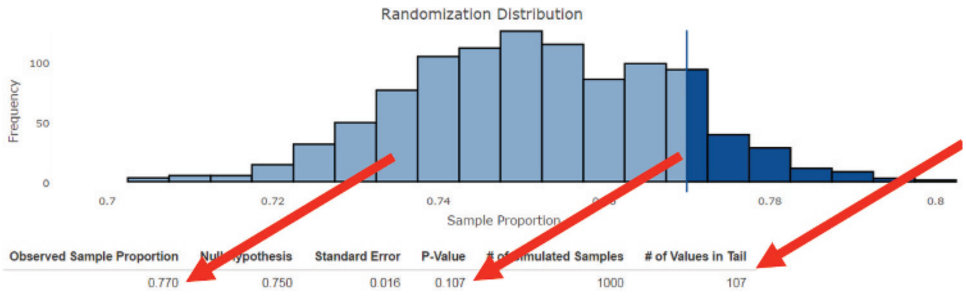
● Upper-Tailed (>)

○ Lower-Tailed (<)

Select number of simulated samples to generate:

○ 1 ○ 10 ○ 100 ● 1,000 ○ 10,000

Copyright 2025 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s). Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.



Observed Sample Proportion	Null Hypothesis	Standard Error	P-Value	# of Simulated Samples	# of Values in Tail
0.770	0.750	0.016	0.107	1000	107

The approximate P-value is 0.107. Because the P-value is greater than 0.05, you would fail to reject H_0. The sample does not provide convincing evidence that more than 75% of Canadians age 18 to 24 would respond "yes" when asked if the statement "When nothing is occupying my attention the first thing I do is reach for my phone" describes them. This is consistent with the P-value obtained for the large-sample z test in Example 10.11 (which was 0.092), and so the same conclusion is reached using either method to approximate the P-value. This is not surprising because the sample size is large enough for the large-sample test to be appropriate. The advantage of the randomization test is that it can be used even when the sample size is not large.

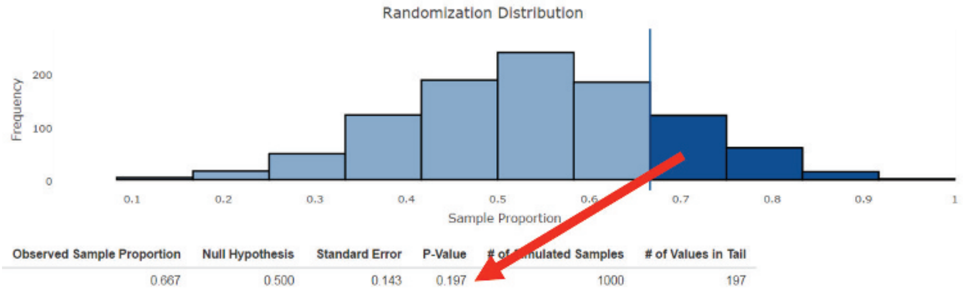
Example 10.21 Vaccination Coverage

Following a 2-year period when the Australian government offered women age 13 to 26 free vaccinations to protect against cervical cancer, researchers carried out surveys of random samples of women in each of the territories in Australia **("Human Papillomavirus (HPV) Vaccination Coverage in Young Australian Women Is Higher Than Previously Estimated," *Vaccine* [2014]: 592–597)**. While the sample sizes were large for most territories, for the Northern Territory the sample size was only 12. The researchers found that 8 of the 12 women in the sample from the Northern Territory had received at least one of the three recommended doses of HPV vaccine. The data in this sample can be used to test the hypothesis that more than half of women age 13 to 26 in the Northern Territory have received at least one dose of the HPV vaccine.

Let p represent the proportion of women age 13 to 26 in the Northern Territory who have received at least one dose of HPV vaccine. The hypotheses of interest are H_0: $p = 0.50$ and H_a: $p > 0.50$. The observed value of the sample proportion is $\hat{p} = 8/12 = 0.667$.

Notice that the large-sample z test of Section 10.3 is not an appropriate choice because the sample size condition of $np \geq 10$ is not met ($np = (12)(0.50) = 6 < 10$). However, because the sample was a random sample and the randomization test does not require a large sample, it is appropriate to proceed with a randomization test.

From the accompanying output from the Shiny app "Randomization Test for One Proportion," the approximate P-value for this upper-tailed randomization test is 0.197. This means that 197 of the 1000 simulated sample proportions were at least as large as 0.667. Using a significance level of 0.05, the null hypothesis would not be rejected. The sample does not provide convincing evidence that more than half of women age 13 to 26 in the Northern Territory of Australia have received at least one dose of HPV vaccine.



Observed Sample Proportion	Null Hypothesis	Standard Error	P-Value	# of Simulated Samples	# of Values in Tail
0.667	0.500	0.143	0.197	1000	197

Example 10.21 illustrates that randomization tests may be used even for small random samples when the conditions for large-sample inference may not be met. For this example, if a large-sample *z* test had been used, then the resulting *P*-value of 0.124 would be quite different than the *P*-value from the randomization test. This illustrates why a hypothesis test procedure should not be used when its assumptions are not met.

An Exact Binomial Test for One Proportion

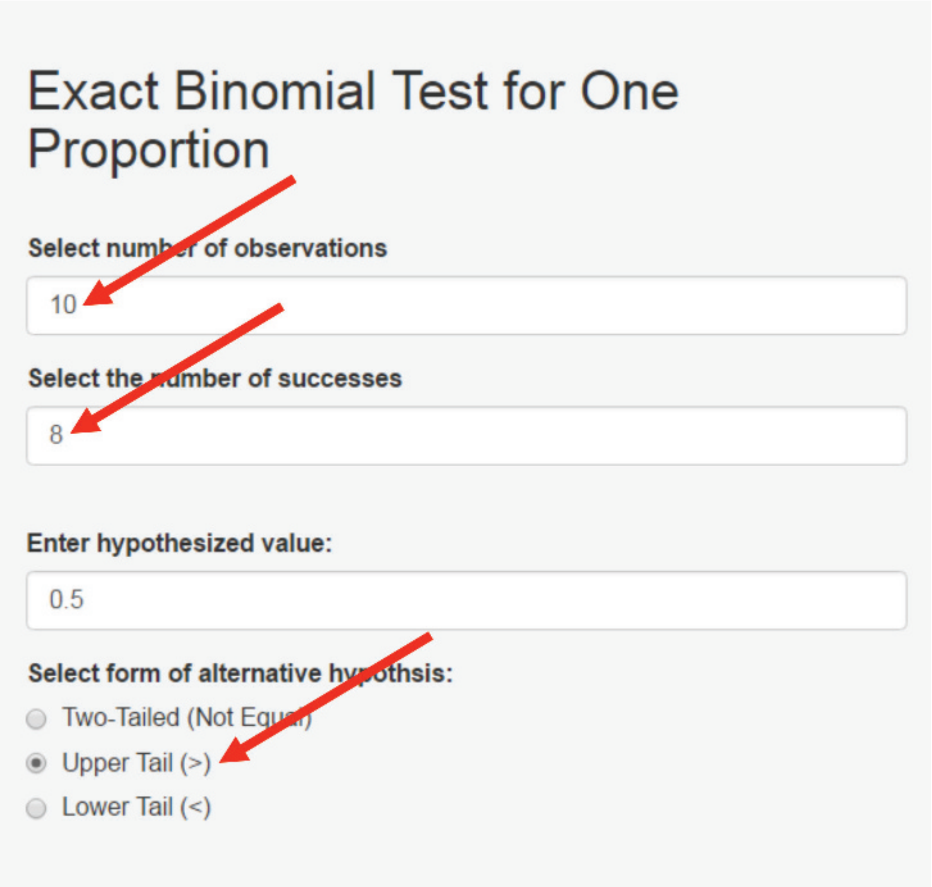
Another way to obtain a *P*-value when testing hypotheses about a population proportion is to use an exact probability approach that uses the binomial distribution. The binomial probability distribution was introduced in Section 7.5, Binomial and Geometric Distributions.

Example 10.22 An Exact Binomial Test

Consider a test of the null hypothesis H_0: $p = 0.50$. The value of *p* that is specified in the null hypothesis, when combined with an observed sample size, identifies a specific binomial distribution that can be used to calculate an "exact" *P*-value. Suppose that the alternative hypothesis of interest is H_a: $p > 0.50$ and that the test will be carried out using data from a random sample of $n = 10$ independent success (S) or failure (F) observations.

Because the sample size is small, the large-sample *z* test is not an appropriate way to test these hypotheses. The exact binomial test does not require a large sample, so it can be used when the sample size is small and the sampling distribution of $\hat{p}$ may not be approximately normal.

Suppose that you observe $x = 8$ successes in the sample of size $n = 10$. This means that $\hat{p} = 8/10 = 0.8$. The binomial distribution with $n = 10$ and $p = 0.50$ (the hypothesized proportion) can be used to calculate the probability of observing a sample proportion as or more extreme than what was observed in the sample. This is the *P*-value for the hypothesis test.



Exact Binomial Test for One Proportion

Select number of observations

10

Select the number of successes

8

Enter hypothesized value:

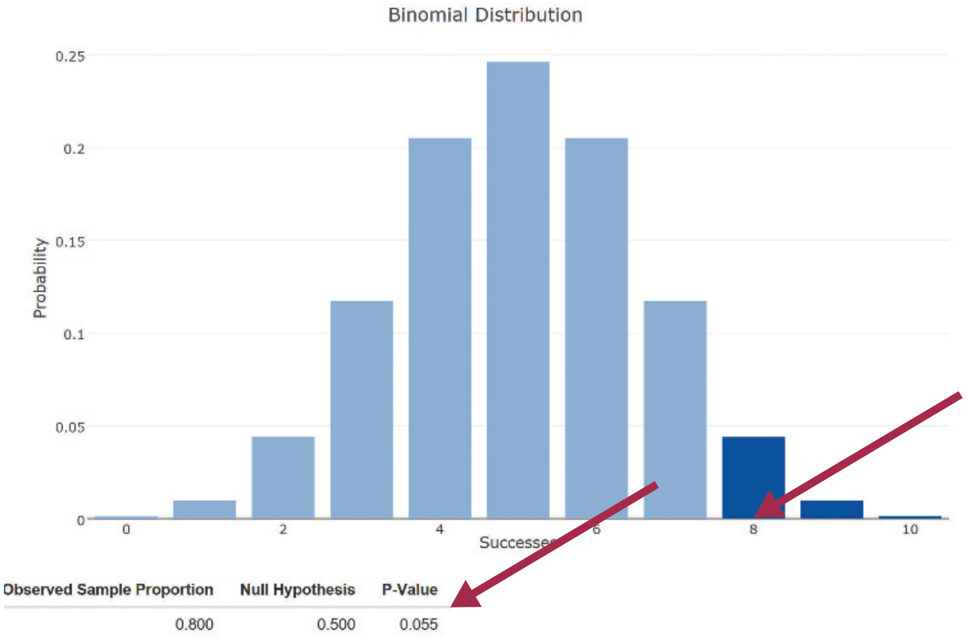
0.5

Select form of alternative hypothsis:

○ Two-Tailed (Not Equal)
● Upper Tail (>)
○ Lower Tail (<)

To calculate the P-value for an exact binomial test, use the Shiny app "Exact Binomial Test for One Proportion." This is one of the Shiny web apps that accompany this text. These web apps are located at **www.cengage.com/Peck**. Enter 10 for the Number of observations and 8 for the Number of successes. The default hypothesized value is already 0.5. Specify "Upper Tail (>)" for the form of the alternative hypothesis. (See the figure at the bottom of the page.).

The Shiny app automatically updates the probability histogram to reflect your choices and displays the upper-tail P-value as a binomial probability. In this example, the P-value is equal to 0.055.



Observed Sample Proportion	Null Hypothesis	P-Value
0.800	0.500	0.055

The probability $P(x \geq 8)$ represents the exact one-sided P-value for testing H_0: $p = 0.50$ versus H_a: $p > 0.50$. From the Shiny app, this probability is 0.055, which is the value of the P-value for the exact binomial test.

Since this P-value $= 0.055$ is greater than $\alpha = 0.05$, using a significance level of 0.05 you would fail to reject H_0. Observing eight successes in $n = 10$ trials is not convincing evidence that $p > 0.50$.

The following examples illustrate the use of the exact binomial test. By revisiting the examples previously used to illustrate the randomization test, comparisons can be made between the approximate P-value from the randomization test and the P-value from the exact binomial test.

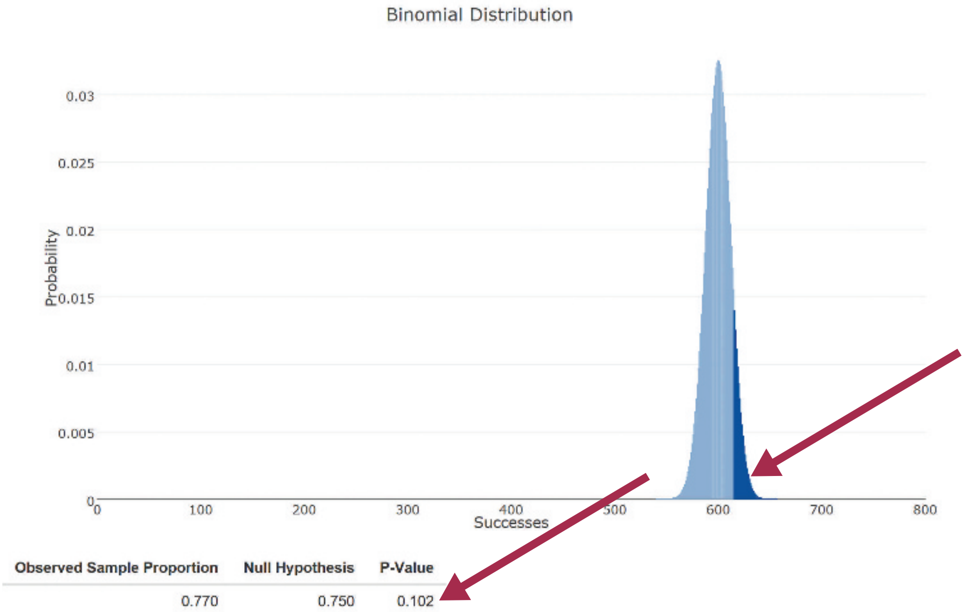
Example 10.23 Cell Phones Revisited Again

In the study described in Examples 10.11 and 10.20, 77% of 800 Canadians age 18 to 24 years responded "yes" when asked if the statement "When nothing is occupying my attention the first thing I do is reach for my phone" describes them. As was done in the previous examples, this data will be used to decide whether the sample supports the claim that more than 75% of Canadians in this age group would respond "yes." In this example, an exact binomial test will be used.

For the exact binomial test, the assumption is made that the Canadians sampled are representative of the population of interest, but there are no sample size conditions that must be satisfied.

The Shiny app "Exact Binomial Test for One Proportion" provides the binomial probability that at least 616 of the people in a random sample of size 800 would have responded "yes" if the null hypothesis of $p = 0.75$ were true.

Copyright 2025 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s). Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.



Observed Sample Proportion	Null Hypothesis	P-Value
0.770	0.750	0.102

From the Shiny app, the P-value $= P(x \geq 616$ when $n = 800$ and H_0: $p = 0.75$ is true) is 0.102. This is greater than 0.05, so you would fail to reject H_0. The sample does not provide convincing evidence that more than 75% of Canadians age 18 to 24 would respond "yes" when asked if the statement, "When nothing is occupying my attention the first thing I do is reach for my phone" describes them. The P-value from the exact binomial test is similar to the approximate P-values for the large-sample z test (0.092) and the randomization test (0.107). This is not surprising because any of these three methods are appropriate for this sample size.

Example 10.24 Vaccination Coverage Revisited

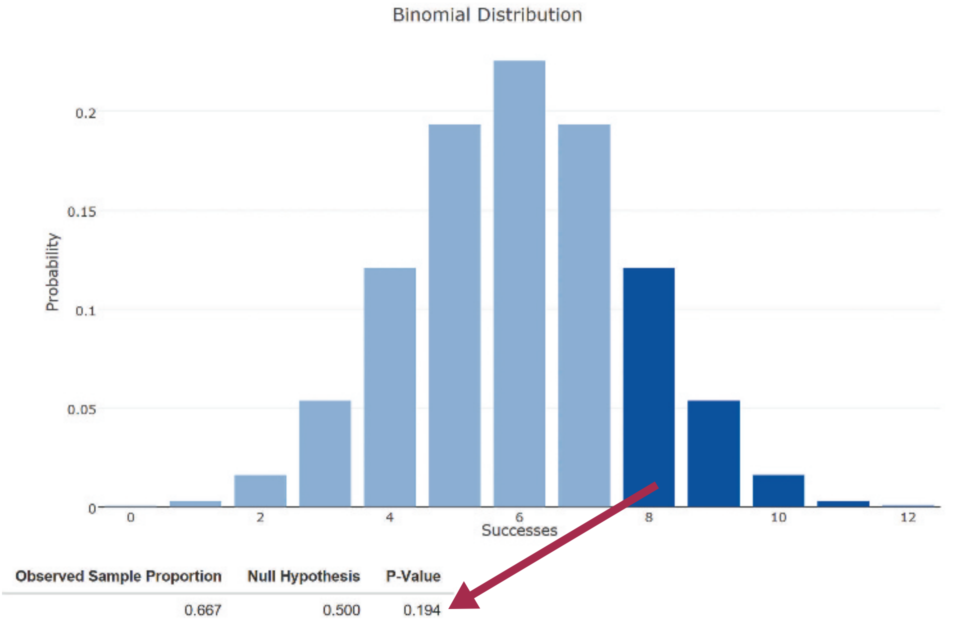
Recall that in Example 10.21, data from a random sample of 12 women from the Northern Territory of Australia were used to test the claim that more than 50% of women age 13 to 26 had received at least one dose of the HPV vaccine. In this example, the exact binomial test will be used to test the hypotheses H_0: $p = 0.5$ and H_a: $p > 0.5$, where p represents the proportion of women age 13 to 26 in the Northern Territory who have had at least one dose of the HPV vaccine. It is reasonable to use the exact binomial test because the sample was a random sample and the exact binomial test does not have any sample size requirements.

Recall that 8 of the 12 women in the sample reported that they had received at least one dose of the vaccine. The binomial distribution with $n = 12$ and $p = 0.5$ (from the null hypothesis) can be used to determine the probability of observing 8 or more successes (which is equivalent to more than 7 successes) in a sample of size 12 when the null hypothesis is true and the actual proportion of successes is 0.5.

From the Shiny app output at the top of the next page, the probability of at least 8 (more than 7) successes in a sample of size 12 when $p = 0.50$ is 0.194. This is the P-value for the exact binomial test, and because the P-value is large, the null hypothesis would not be rejected.

Notice that the P-value for the exact binomial test and the approximate P-value for the randomization test (0.197) are very close, and either method would be an appropriate choice for this sample. The P-value for the large-sample z test was quite a bit smaller (0.124), and you would not want to base a conclusion on this P-value because the assumptions required for the large-sample z test are not met.

Copyright 2025 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s). Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.



Observed Sample Proportion	Null Hypothesis	P-Value
0.667	0.500	0.194

Exercises 10.74–10.79

10.74 Scientists are only beginning to learn about the long-term effects of space travel on human health. A study published in 2016 **(*Nature Scientific Reports* 6, Article number: 29901, nature.com, July 28, 2016)** found that by 2014, seven of the U.S. astronauts who traveled to the moon during Apollo lunar missions of the 1960s and 1970s had died, and that three of these ($3/7 = 43\%$) had died from cardiovascular disease (CVD). The overall U.S. death rate due to CVD for adults age 55 to 64 in 2013 was 27%. Do the data for lunar astronauts indicate that, as a group, they are at increased risk of death caused by CVD? Assume that it is reasonable to regard this sample as representative of the population of past, present, and future U.S. lunar astronauts.

a. Explain why the data in this example should not be analyzed using a large-sample hypothesis test for a population proportion.

b. Use the output at the bottom of the page from the Shiny app "Randomization Test for One Proportion" to complete an appropriate hypothesis test.

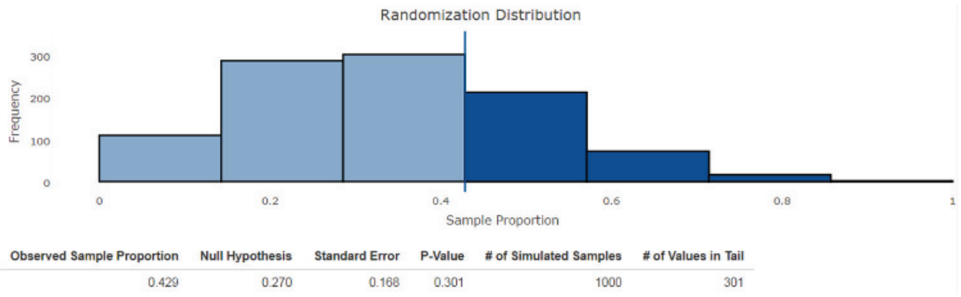
10.75 A study of hospitalized patients who develop pneumonia reported that 1 in 5 (20%) are readmitted to the hospital within 30 days after discharge **("Comparison of Therapist-Directed and Physician-Directed Respiratory Care in COPD Subjects with Acute Pneumonia," *Respiratory Care* [2015]: 151–154)**.

The study reported that 15 out of $n = 162$ hospital patients who had been treated for pneumonia using a respiratory therapist protocol were readmitted to the hospital within 30 days after discharge. These data can be used to decide if the proportion readmitted is less than 0.20.

a. What hypotheses should be tested?

b. Discuss whether the conditions necessary for a large-sample hypothesis test for one proportion are satisfied.

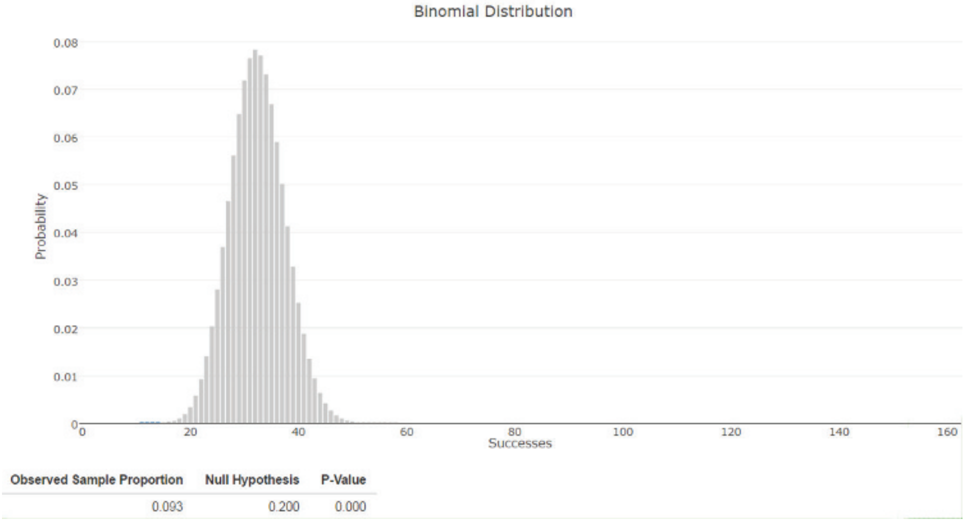
Output for Exercise 10.74



Observed Sample Proportion	Null Hypothesis	Standard Error	P-Value	# of Simulated Samples	# of Values in Tail
0.429	0.270	0.168	0.301	1000	301

Copyright 2025 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s). Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

Output for Exercise 10.75



Observed Sample Proportion	Null Hypothesis	P-Value
0.093	0.200	0.000

c. The exact binomial test can be used even in cases when the sample size condition for the large-sample test is met. Use the output at the top of the page from the Shiny app "Exact Binomial Test for One Proportion" to carry out an exact binomial test for one proportion, using the null hypothesis that the population proportion of subjects who will be readmitted to a hospital within 30 days after following a respiratory therapist protocol for treatment of pneumonia is equal to 1/5 = 0.2.

d. Calculate the *P*-value for the same hypotheses using the large-sample approach for testing one population proportion. Compare this *P*-value to the *P*-value obtained in Part (b). Would the same conclusion be reached in either case? Explain.

10.76 A sample of dogs were trained using a "Do as I do" method, in which the dog observes the trainer performing a simple task (such as climbing onto a chair or touching a chair) and is expected to perform the same task on the command "Do it!" In a separate training session, the same dogs were trained to lie down regardless of the trainer's actions.

Later, the trainer demonstrated a new simple action and said "Do it!" The dog then either repeated the new action, or repeated a previous trained action (such as lying down). The dogs were retested on the new simple action after 1 minute had passed, and after 1 hour had passed. A "success" was recorded if a dog performed the new simple action on the command "Do it!" before performing a previously trained action.

The article **"Your Dog Remembers More Than You Think" (*Science*, November 23, 2016, sciencemag.org)** reports that dogs trained using this method recalled the correct new action in 33 out of 35 trials. Suppose the data from this study will be used to determine if more than half of all dogs trained using this method would recall the correct new action.

a. Explain why the data in this example should not be analyzed using a large-sample hypothesis test for a population proportion.

b. Perform an exact binomial test of the null hypothesis that the proportion of all dogs trained using this method who would perform the correct new action is 0.5, versus the alternative hypothesis that the proportion is greater than 0.5.

10.77 Salmonella bacteria is not uncommon in chicken due to the conditions in which chickens are raised. The article **"Why Ground Meat is so Dangerous" (*Consumer Reports*, August 2022)** states that the USDA permits poultry producers to have salmonella detected in up to 25% of ground chicken samples tested (thus the title of the article!). Consumer Reports tested 75 packages of ground chicken from a variety of producers and found salmonella in 23 of the packages tested.

a. Use the output at the top of the next page from the Shiny app "Randomization Test for One Proportion" to carry out a test to determine if there is convincing evidence that the proportion of ground chicken packages that contain salmonella is greater than the 25% allowed by the USDA. Use a significance level of 0.05 for your test.

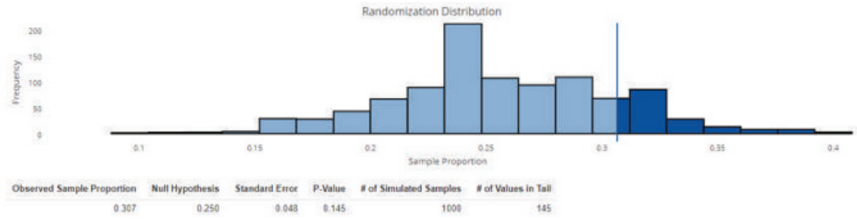
b. What assumption about the sample of 75 packages that were tested would you need to make in order for the randomization test in Part (a) to be appropriate?

c. Is the sample size large enough to satisfy the large-sample requirement of the large-sample z test for a population proportion? If so, carry out the large sample test to determine if there is convincing evidence that the proportion of ground chicken packages that contain salmonella is greater than the 25% allowed by the USDA. Use a significance level of 0.05 for your test.

d. Did you reach the same conclusion in the tests from Part (a) and Part (c). Does this surprise you?

Copyright 2025 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s). Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

Output for Exercise 10.77



Observed Sample Proportion	Null Hypothesis	Standard Error	P-Value	# of Simulated Samples	# of Values in Tail
0.307	0.250	0.048	0.145	1000	145

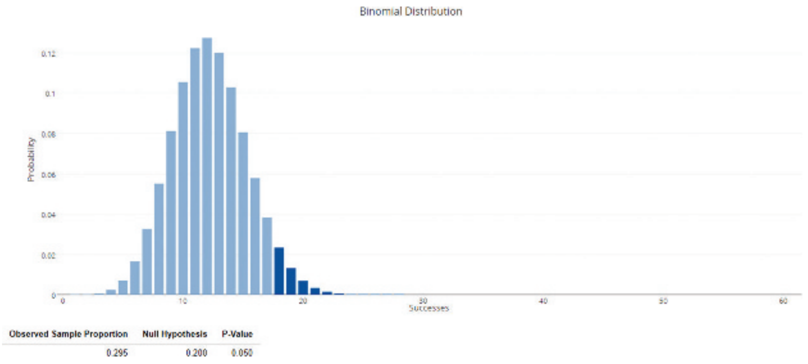
10.78 The study described in Exercise 10.41 surveyed British businesses that experimented with a 4-day work week. In this study, 61 British businesses of various types and sizes gave their workers a paid day off each week for 6 months to see if they could get as much done by working more efficiently when they were at work. At the end of the 6 months, 18 of the companies planned to make the 4-day work week permanent. Suppose that you wanted to carry out a test to determine if more than 20% of all British businesses that implemented a 4-day work week would decide to make the change permanent. Use the given output from the Shiny app "Exact Binominal Test for One Proportion" to carry out the relevant hypothesis test.

10.79 At one point during the 2015 NFL season, head coach Bill Belichick and the New England Patriots had won 19 of their past 25 called coin flips at the beginning of NFL games (**"For Bill Belichick, Patriots' Strategy Is No Flip of the Coin," bostonglobe.com, November 4, 2015**). Suppose that these 25 coin toss calls can be considered as representative of all coin toss calls made by this team.

a. Perform an exact binomial test to determine if there is convincing evidence that the proportion of all coin flip calls that the Patriots win is greater than 0.5.

b. Discuss the conditions required for the exact binomial version of the hypothesis test. Write a brief explanation of why the results of the test performed in Part (a) do not necessarily mean that Coach Belichick is able to predict the results of coin flips better than other coaches.

Output for Exercise 10.78



Observed Sample Proportion	Null Hypothesis	P-Value
0.295	0.200	0.050

Section 10.8 Randomization Test for a Population Mean (Optional)

The randomization approach introduced in Section 10.7 can also be used to approximate *P*-values for tests of hypotheses about a population mean.

Example 10.25 Getting Enough Sleep (Revisited)

● Example 10.13 gave data on reported sleep time (in minutes) on nights where there is school the next day for a representative sample of 10 students that are consistent with summary quantities given in the paper **"Exposure to Electronic Devices and Sleep Quality in Adolescents: A Matter of Type, Duration, and Timing" (*Sleep Health* [2020]: 172–178)**. The original sample data values are given in the table below.

Student ID	1	2	3	4	5	6	7	8	9	10
Sleep Time	396	436	573	491	434	545	490	345	490	401

● **Data set available online**

The sample mean calculated from the original sample is $\bar{x} = 460.1$ minutes. The question posed in Example 10.13 was whether there was evidence that the mean sleep time for students on nights when they have school the next day is less than the minimum recommendation of 8 hours (480 minutes). To answer this question, the following hypotheses were tested:

$$H_0\colon \mu = 480 \qquad H_0\colon \mu < 480$$

The distance between the hypothesized mean of 480 minutes (from the null hypothesis) and the sample mean from the original sample is $480 - 460.1 = 19.9$ minutes.

The data values in the original sample can be shifted to represent a sample from a hypothetical population with mean $\mu = 480$ minutes by adding 19.9 minutes to each value, as shown here:

Student ID	1	2	3	4	5	6	7	8	9	10
Sleep Time	396	436	573	491	434	545	490	345	490	401
Sleep Time + 19.9	415.9	455.9	592.9	510.9	453.9	564.9	509.9	364.9	509.9	420.9

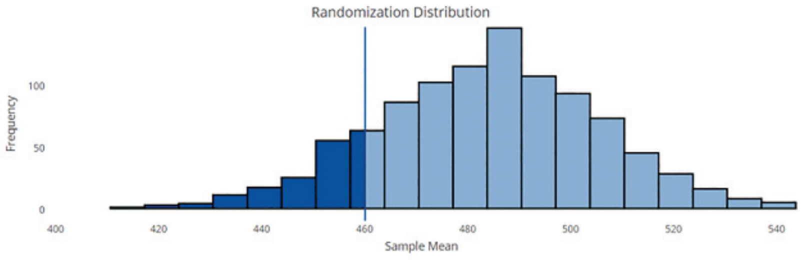
A randomization approach based on resampling can be used in this example, instead of a one-sample *t* test. The sample was a representative sample of students, and this is the only condition needed for the randomization test.

Begin by resampling from the sample that has been shifted to have a mean of 480 minutes. That is, you can draw new simulated samples, selected at random with replacement from the shifted sample. Here is one sample from the shifted sample:

Student ID	9	2	5	9	3	2	2	2	1	4
Sleep Time + 19.9	509.9	455.9	453.9	509.9	592.9	455.9	455.9	455.9	415.9	510.9

The sample mean for this sample is $\bar{x} = 481.7$ minutes.

The randomization distribution is generated by selecting many random samples from the shifted sample. This distribution can be viewed as an approximate sampling distribution for $\bar{x}$ under the assumption that the null hypothesis $\mu = 480$ is true. A *P*-value is the probability of obtaining a test statistic value at least as inconsistent with H_0 as what actually resulted. So, the *P*-value for testing the null hypothesis $H_0\colon \mu = 480$ minutes against the one-sided alternative $H_a\colon \mu < 480$ minutes is the probability of observing $\bar{x} \le 460.1$ when H_0 is true. This probability can be approximated using the proportion of simulated values of $\bar{x}$ that fall at or below 460.1 in the randomization distribution. Below is the output from the Shiny App "Randomization Test for One Mean." This Shiny app can be found in the app collection at **www.cengage.com/Peck**.



Observed Sample Mean	Simulation Standard Error	Lower Tailed P-value	Number of Values in Tail	Number of Simulated Samples
460.100	21.094	0.166	166	1000

The randomization test one-sided *P*-value is *P*-value = 166/1000 = 0.166. This *P*-value is greater than $\alpha = 0.05$, so you fail to reject H_0. The *P*-value obtained from the one-sample *t* test in Example 10.13 was 0.198, so in this case, the conclusion in the hypothesis test does not change regardless of which of these two methods is used to carry out the test.

Because the null hypothesis was not rejected, you would conclude that there is not convincing evidence that the mean sleep time on nights when there is school the next night is less than 480 minutes (8 hours).

Exercises 10.80–10.85

● **Data set available online**

10.80 ● The dodo was a species of flightless bird that lived on the island of Mauritius in the Indian Ocean. The first record of human interaction with the dodo occurred in 1598, and within 100 years the dodo was extinct due to hunting by humans and other newly introduced invasive species. After the extinction, the word "dodo" became synonymous with stupidity, implying that the birds lacked the intelligence to avoid or escape extinction. The closest existing relatives of the dodo are pigeons and doves.

Researchers at the American Museum of Natural History used computed tomography scans to measure the brain size ("endocranial capacity") of one of the few existing preserved dodo birds, and to measure the brain sizes (reported in log mm^3) in samples of eight birds that are close relatives of dodos **("The First Endocast of the Extinct Dodo and an Anatomical Comparison Amongst Close Relatives," *Zoological Journal of the Linnean Society* [2016]: 950–953)**.

The brain size for the dodo was 4.17 log mm^3. The following table contains the brain sizes for the sample of birds from related species (approximate values from a graph in the paper).

Brain Size	Brain Size
3.08	3.52
3.16	3.78
3.33	4.00
3.35	

a. Use the Shiny app output at the bottom of the page to carry out a randomization test of the hypothesis that the population mean brain size for birds that are relatives of the dodo differs from the established dodo brain size of 4.17.

b. What does the result of the test indicate about the brain size of the dodo?

10.81 ● Just how dirty is the money you carry in your wallet? The authors of the article **"Dirty Money: New Research Reveals the Filthiness of Our Cash, Cards, & Coins" (lendedu.com, 2020)** measured the bacteria on the surface of credit and debit cards, bills, and coins. Germ scores (with higher scores corresponding to dirtier surfaces) for 12 coins of various denominations, arranged in order from greatest germ score to lowest, are given. You may consider these germs scores to be a representative sample of germs scores for coins.

356 329 228 227 109 92
84 60 52 33 32 25

a. Construct a graphical display for the data. What characteristic of the distribution indicates that the methods based on the *t* distribution may not be appropriate?

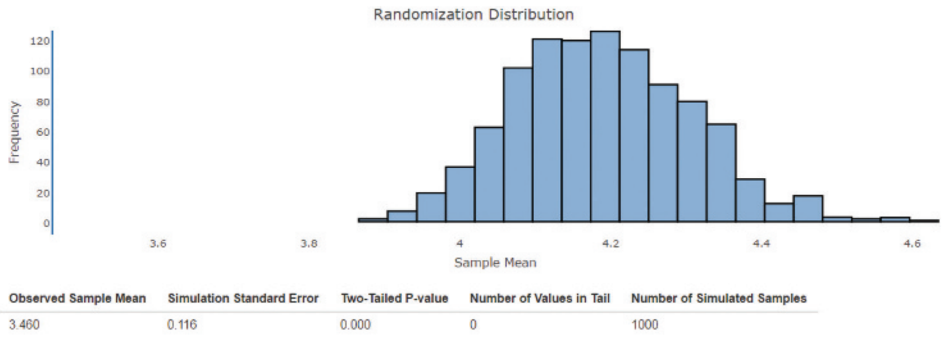
b. Use a randomization test (Shiny app: "Randomization Test for One Mean") to perform a test of the hypothesis that the population mean germ score for coins is greater than 50. (For reference, it is recommended that a food preparation surface have a germ score of 10 or less.) Use a 0.05 significance level.

10.82 ● The paper referenced in the previous exercise also gave data on germs scores for bills of various denominations. You may consider these germs scores to be a representative sample of germs scores for bills.

633 609 349 273 259 242 200 163 143 129 126
119 119 108 105 104 98 78 74 68 56 55
54 41 39 35 32

a. What characteristic of this sample of germ scores indicates that the methods based on the *t* distribution may not be appropriate?

Output for Exercise 10.80



Observed Sample Mean	Simulation Standard Error	Two-Tailed P-value	Number of Values in Tail	Number of Simulated Samples
3.460	0.116	0.000	0	1000

Copyright 2025 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s). Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

b. Use a randomization test (Shiny app: "Randomization Test for One Mean") to perform a test of the hypothesis that the population mean germ score for bills is greater than 150. (For reference, it is recommended that a food preparation surface have a germ score of 10 or less.)

10.83 ● Researchers studied ergometer (rowing machine) times for international competitors in the 2007 World Junior Rowing Championships. They found that the mean time to row 2000 meters on an ergometer for the population of international sculls competitors was 387 seconds (**"Does 2000-m Rowing Ergometer Performance Time Correlate with Final Rankings at the World Junior Rowing Championship? A Case Study of 398 Elite Junior Rowers,"** ***Journal of Sports Sciences*** **[2009]: 361–366)**.

A related research team studied a sample of 24 junior sculls rowers from only the United States in 2013–2014 and reported summary statistics for their 2000-meter ergometer times (**"Correlates of Performance at the U.S. Rowing Youth National Championships: A Case Study of 152 Junior Rowers,"** ***The Sport Journal*****, March 3, 2014)**. Data for a representative sample of 24 rowers that are consistent with summary statistics given in the paper are shown in the accompanying table.

Rower ID	Time (seconds)	Rower ID	Time (seconds)
1	391	13	404
2	389	14	399
3	386	15	390
4	396	16	395
5	395	17	392
6	409	18	401
7	406	19	381
8	402	20	392
9	393	21	376
10	392	22	400
11	382	23	388
12	401	24	402

a. Use the Shiny app output at the bottom of the page to carry out a randomization test of the hypothesis that the population mean 2000-meter ergometer time for U.S. junior sculls rowers differs from the 2007 international standard of 387 seconds. Use a significance level of 0.05.

b. Based on the result of the hypothesis test, does it seem that the U.S. junior sculls rowers have "caught up," on average, with the international championship rowers from 2007? Explain.

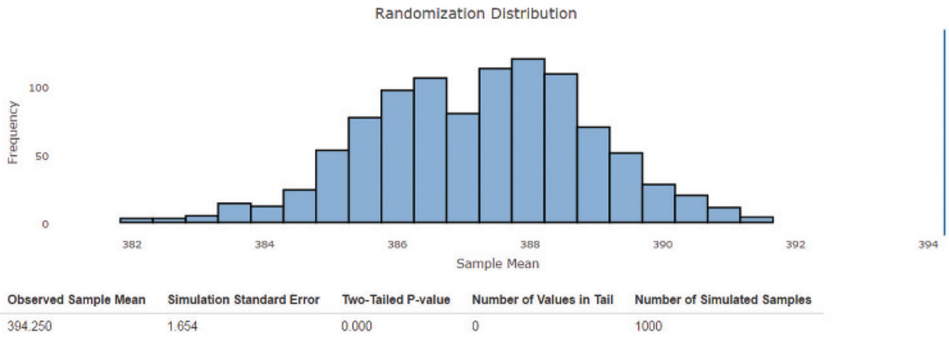
10.84 ● In their senior thesis, students at Pomona College posed questions about how stocks with clever ticker symbols (like PZZA, the ticker for Papa John's Pizza, or BOOM, the ticker for Explosive Fabricators) performed relative to the stock market in general (**"The Name Game: The Importance of Resourcefulness, Ruses, and Recall in Stock Ticker Symbols," 2019)**. This raises an interesting question that can be considered using the data in the table at the top of the next page, which represent the total 1-year percent return as of March 1, 2022, for a sample of 12 stocks with clever ticker symbols. For purposes of this exercise, you can assume that this sample is representative of the population of stocks with clever ticker symbols.

a. Construct a dotplot of the percent return data. What about this plot suggests that methods based on the *t* distribution may not be appropriate?

b. Use the output at the top of the next page page from the Shiny App "Randomization Test for One Mean" to test the hypothesis that the population mean percent annual return for stocks with clever ticker symbols is less than 6.06%, the percent return for the stocks that are part of the Dow Jones stock index for the same period of time.

10.85 ● Major League Baseball (MLB) includes two groups of teams, in two "leagues." There are 15 teams in each of the American League (AL) and the National League (NL). Since 1997, teams in each of the leagues play teams from the other league in "interleague" regular-season games.

Output for Exercise 10.83

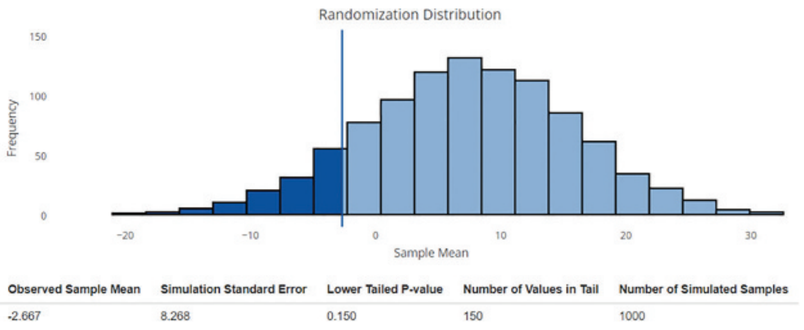


```
Observed Sample Mean   Simulation Standard Error   Two-Tailed P-value   Number of Values in Tail   Number of Simulated Samples
394.250                1.654                       0.000                0                          1000
```

Copyright 2025 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s). Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

Table and Output for Exercise 10.84

Ticker	BOOM	BUD	CASH	CAKE	FUN	LUV	PZZA	ROCK	SHOO	WOOF	YUM	ZEUS
One-Year Return	−53.3	8.0	18.8	−26.6	15.4	24.7	17.3	−46.6	14.0	−14.7	20.4	40.0



Observed Sample Mean	Simulation Standard Error	Lower Tailed P-value	Number of Values in Tail	Number of Simulated Samples
-2.667	8.268	0.150	150	1000

One way to determine whether one league is stronger than the other is to consider the interleague winning percentages for all the teams in one of the two leagues, say, the National League, for one season. For purposes of this exercise, consider the interleague games played in the 2022 season to be a representative sample of the performance of the teams in a population of potential future seasons.

Here are the 2022 interleague winning percentages for the 15 NL teams:

Team	**W**	**L**	**Winning Percentage**
ARI	12	8	60
ATL	13	7	65
CHC	6	14	30
CIN	12	8	60
COL	9	11	45
LAD	15	5	75
MIA	8	12	40
MIL	15	5	75
NYM	9	11	45
PHI	9	11	45
PIT	4	16	20
SD	8	12	40
SF	10	10	50
STL	10	10	50
WSH	8	12	40

(continued)

a. What characteristic of this sample of NL team interleague winning percentages indicates that the methods based on the *t* distribution may not be appropriate?

b. Use a randomization test (Shiny app: "Randomization Test for One Mean") to perform a test of the hypothesis that the population mean interleague winning percentage for NL teams differs from 50%.

c. Use the results from Part (b) to explain whether it is reasonable to say that the National League or the American League performs significantly better than the other in interleague play.

Chapter Activities

Activity 10.1 Comparing the *t* and *z* Distributions Technology Activity

This activity requires use of a computer or a graphing calculator.

The purpose of this activity is to use simulation from a known normal population to investigate how the behavior of

$$t = \frac{\bar{x} - \mu}{\frac{s}{\sqrt{n}}} \text{ compares to the behavior of } z = \frac{\bar{x} - \mu}{\frac{\sigma}{\sqrt{n}}}.$$

The instructions that follow assume the use of Minitab. If you are using a different software package or a graphing calculator, your instructor will provide alternative instructions.

Background: Suppose a random sample will be selected from a population that is known to have a normal distribution. Then the statistic

$$z = \frac{\bar{x} - \mu}{\frac{\sigma}{\sqrt{n}}}$$

Copyright 2025 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s). Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

has a standard normal (z) distribution. Since it is rarely the case that σ is known, inferences for population means are usually based on the statistic $t = \frac{\bar{x} - \mu}{(s/\sqrt{n})}$, which has a t distribution rather than a z distribution. The informal justification for this was that the use of s to estimate σ introduces additional variability, resulting in a statistic whose distribution is more spread out than is the z distribution.

In this activity, you will use simulation to sample from a known normal population and then investigate how the behavior of $t = \frac{\bar{x} - \mu}{s/(\sqrt{n})}$ compares with the behavior of $z = \frac{\bar{x} - \mu}{\sigma/(\sqrt{n})}$.

1. Generate 200 random samples of size 5 from a normal population with mean 100 and standard deviation 10.

Using Minitab, go to the Calc Menu. Then

Calc→ Random Data → Normal
In the "Generate" box, enter 200
In the "Store in columns" box, enter c1-c5
In the mean box, enter 100
In the standard deviation box, enter 10
Click on OK

You should now see 200 rows of data in each of the first 5 columns of the Minitab worksheet.

2. Each row contains five values that have been randomly selected from a normal population with mean 100 and standard deviation 10. Viewing each row as a sample of size 5 from this population, calculate the mean and standard deviation for each of the 200 samples (the 200 rows) by using Minitab's row statistics functions, which can also be found under the Calc menu:

Calc→ Row statistics
Choose the "Mean" button
In the "Input Variables" box, enter c1-c5
In the "Store result in" box, enter c7
Click on OK

You should now see the 200 sample means in column 7 of the Minitab worksheet. Name this column "x-bar" by typing the name in the gray box at the top of c7.

Now follow a similar process to compute the 200 sample standard deviations, and store them in c8. Name c8 "s."

3. Next, calculate the value of the z statistic for each of the 200 samples. We can calculate z in this example because we know that the samples were selected from a population for which $\sigma = 10$.

Use the calculator function of Minitab to compute $z = \frac{\bar{x} - \mu}{(\sigma/\sqrt{n})} = \frac{\bar{x} - 100}{(10/\sqrt{5})}$ as follows:

Calc → Calculator
In the "Store results in" box, enter c10
In the "Expression box" type in the following: (c7-100)/(10/sqrt(5))
Click on OK

You should now see the z values for the 200 samples in c11. Name c11 "z."

4. Now calculate the value of the t statistic for each of the 200 samples. Use the calculator function of Minitab to compute $t = \frac{\bar{x} - \mu}{(s/\sqrt{n})} = \frac{\bar{x} - 100}{(s/\sqrt{5})}$ as follows:

Calc → Calculator
In the "Store results in" box, enter c11
In the "Expression box" type in the following: (c7-100)/(c8/sqrt(5))
Click on OK

You should now see the t values for the 200 samples in c10. Name c10 "t."

5. Graphs, at last! Now construct histograms of the 200 z values and the 200 t values. These two graphical displays will provide insight about how each of these two statistics behaves in repeated sampling. Use the same scale for the two histograms so that it will be easier to compare the two distributions.

Graph → Histogram
In the "Graph variables" box, enter c10 for graph 1 and c11 for graph 2
Click the Frame dropdown menu and select multiple graphs.
Then under the scale choices, select "Same X and same Y."

6. Now use the histograms from Step 5 to answer the following questions:

a. Write a brief description of the shape, center, and spread for the histogram of the z values. Is what you see in the histogram consistent with what you would have expected to see? Explain. (Hint: In theory, what is the distribution of the z statistic?)

b. How does the histogram of the t values compare to the z histogram? Be sure to comment on center, shape, and spread.

c. Is your answer to Part (b) consistent with what would be expected for a statistic that has a t distribution? Explain.

d. The z and t histograms are based on only 200 samples, and they only approximate the corresponding sampling distributions. The 5th percentile

Copyright 2025 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s). Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

for the standard normal distribution is -1.645 and the 95th percentile is $+1.645$. For a t distribution with $\text{df} = 5 - 1 = 4$, the 5th and 95th percentiles are -2.13 and $+2.13$, respectively. How do these percentiles compare to those of the distributions displayed in the histograms? (Hint: Sort the 200 z values—in Minitab, choose “Sort” from the Manip menu. Once the values are sorted, percentiles from the histogram can be found by counting in 10 [which is 5% of 200] values from either end of the sorted list. Then repeat this with the t values.)

e. Are the results of your simulation and analysis consistent with the statement that the statistic $z = \dfrac{\bar{x} - \mu}{(\sigma/\sqrt{n})}$ has a standard normal (z) distribution and the statistic $t = \dfrac{\bar{x} - \mu}{(s/\sqrt{n})}$ has a t distribution? Explain.

Activity 10.2 A Meaningful Paragraph

The purpose of this activity is to check your understanding of some of the concepts and terminology introduced in this chapter.

Write a meaningful paragraph that includes the following six terms: **hypotheses, *P*-value, reject H_0, Type I error, statistical significance, practical significance.**

A “meaningful paragraph” is a coherent piece of writing in an appropriate context that uses all of the listed words. The paragraph should show that you understand the meaning of the terms and their relationship to one another. A sequence of sentences that just define the terms is *not* a meaningful paragraph. When choosing a context, think carefully about the terms you need to use. Choosing a good context will make writing a meaningful paragraph easier.

Summary Key Concepts and Formulas

Term or Formula	Comment
Hypothesis	A claim about the value of a population characteristic.
Null hypothesis, H_0	The hypothesis initially assumed to be true. It has the form H_0: population characteristic = hypothesized value.
Alternative hypothesis, H_a	A hypothesis that specifies a claim that is contradictory to H_0 and is judged the more plausible claim when H_0 is rejected.
Type I error	Rejecting H_0 when H_0 is true. The probability of a Type I error is denoted by α and is referred to as the significance level for the test.
Type II error	Not rejecting H_0 when H_0 is false. The probability of a Type II error is denoted by β.
Test statistic	A value calculated from sample data that is then used as the basis for making a decision between H_0 and H_a.
***P*-value**	The probability, computed assuming H_0 is true, of obtaining a value of the test statistic at least as contradictory to H_0 as what actually resulted. H_0 is rejected if P-value $\le \alpha$ and not rejected if P-value $> \alpha$, where α is the chosen significance level.

Term or Formula	Comment
$z = \dfrac{\hat{p} - \text{hypothesized value}}{\sqrt{\dfrac{(\text{hyp val})(1 - \text{hyp val})}{n}}}$	A test statistic for testing H_0: p = hypothesized value when the sample size is large. The P-value is determined as an area under the z curve.
$z = \dfrac{\bar{x} - \text{hypothesized value}}{\dfrac{\sigma}{\sqrt{n}}}$	A test statistic for testing H_0: μ = hypothesized value when σ is known and either the population distribution is normal or the sample size is large. The P-value is determined as an area under the z curve.
$t = \dfrac{\bar{x} - \text{hypothesized value}}{\dfrac{s}{\sqrt{n}}}$	A test statistic for testing H_0: μ = hypothesized value when σ is unknown and either the population distribution is normal or the sample size is large. The P-value is determined as an area under the t curve with $\text{df} = n - 1$.
Power	The power of a test is the probability of rejecting the null hypothesis. Power is affected by the size of the difference between the hypothesized value and the actual value, the sample size, and the significance level.

Copyright 2025 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s). Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

Chapter Review Exercises 10.86–10.100

● **Data set available online**

10.86 The report **"A Crisis in Civic Education" (American Council of Trustees and Alumni, January 2016, goacta.org)** summarizes data from a survey of a representative sample of 1000 adult Americans regarding their understanding of U.S. government. Only 459 of the adults in the sample were able to give a correct response to a question asking them to choose a correct definition of the Bill of Rights from a list of five possible answers. Using a significance level of 0.01, determine if there is convincing evidence that less than half of adult Americans could identify the correct definition of the Bill of Rights.

10.87 A study of sleeping habits estimated that 52% of adult Americans grew up sleeping with a security blanket or stuffed animal **("Bedtime Habits Survey," studyfinds.org, June 10, 2023)**. This estimate was based on a sample of 2000 adult Americans. For purposes of this exercise, you can assume that the sample is representative of the population of adult Americans.

a. Use the given information to construct and interpret a 95% confidence interval for the proportion of adult Americans who grew up sleeping with a security blanket or stuffed animal.

b. Is 0.5 included in your confidence interval from Part (a)? Based on your confidence interval, is it reasonable to think that a majority of adult Americans grew up sleeping with a security blanket or stuffed animal?

10.88 The study described in the previous exercise found that of the 1040 people surveyed who said that they grew up sleeping with a security blanket or stuffed animal, 801 said that they still do, even when sharing a bed with a partner. For purposes of this exercise, you can assume that the sample is representative of the population of adult Americans who grew up sleeping with a security blanket or stuffed animal. Do the data from this study provide convincing evidence that the proportion of adult Americans who grew up sleeping with a security blanket or stuffed animal who still do as an adult is greater than 0.50? Test the relevant hypotheses using a significance level of $\alpha = 0.01$.

10.89 The authors of the paper **"Mean Platelet Volume Could Be Possible Biomarker in Early Diagnosis and Monitoring of Gastric Cancer" (*Platelets* [2014]: 592–594)** wondered if mean platelet volume (MPV) might be a way to distinguish patients with gastric cancer from patients who did not have gastric cancer. MPV was recorded for 31 patients with gastric cancer. The sample mean was reported to be 8.31 femtoliters (fL) and the sample standard deviation was reported to be 0.78 fL. For a group of healthy people, the mean MPV was 7.85 fL. Is there convincing evidence that the mean MPV for patients with gastric cancer is greater than 7.85 fL? For purposes of this exercise, you can assume that the sample of 31 patients with gastric cancer is representative of the population of all patients with gastric cancer.

10.90 People in a random sample of 236 students enrolled at a liberal arts college were asked questions about how many hours of sleep they get each night **("Alcohol Consumption, Sleep, and Academic Performance Among College Students," *Journal of Studies on Alcohol and Drugs* [2009]: 355–363)**. The sample mean sleep duration (average hours of daily sleep) was 7.71 hours and the sample standard deviation was 1.03 hours. The recommended number of hours of sleep for college-age students is 8.4 hours. Is there convincing evidence that the population mean sleep duration for students at this college is less than the recommended number of 8.4 hours? Test the relevant hypotheses using $\alpha = 0.01$.

10.91 *USA TODAY* reported that Americans spend 4.1 hours per weekday checking work email (October 14, 2016). This was an estimate based on a survey of 1004 white-collar workers in the United States.

a. Suppose that you would like to know if there is evidence that the mean time spent checking work email for white-collar workers in the United States is more than half of the 8-hour workday. What would you need to assume about the sample to use the given sample data to answer this question?

b. Given that any concerns about the sample were satisfactorily addressed, carry out a test to decide if there is evidence that the mean time spent checking work email for white-collar workers in the United States is more than half of the 8-hour workday.

10.92 According to a large national survey conducted by the Pew Research Center **("What Americans Think About NSA Surveillance, National Security and Privacy," May 2, 2015, pewresearch.org)**, 54% of adult Americans disapprove of the National Security Agency collecting records of phone and Internet data. Suppose that this estimate was based on a random sample of 1000 adult Americans.

a. Is there convincing evidence that a majority of adult Americans feel this way? Test the relevant hypotheses using a 0.05 significance level.

b. The actual sample size was much larger than 1000. If you had used the actual sample size when

Copyright 2025 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s). Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

doing the calculations for the test in Part (a), would the *P*-value have been larger than, the same as, or smaller than the *P*-value you obtained in Part (a)? Provide a justification for your answer.

10.93 In a sample of 1537 adult Americans, 69% of those surveyed indicated that they would be willing to wear a fitness tracker if it would give them a discount on their health insurance **("Nearly 70% of Americans Would Wear a Fitness Tracker /Smartwatch for Discounted Health Insurance," valuepenguin.com, April 18, 2022)**. You may assume that this sample is representative of the population of adult Americans. Use the given information to construct and interpret a 99% confidence interval for the proportion of adult Americans who would be willing to wear a fitness tracker if it would give them a discount on their health insurance.

10.94 The article **"Facebook Use and Academic Performance Among College Students" (*Computers in Human Behavior* [2015]: 265–272)** estimated that 87% percent of students at a large public university in California who are Facebook users update their status at least two times a day. This estimate was based on a random sample of size 261. Assume that this sample is representative of the students at this university.

a. Does this sample provide convincing evidence that more than 80% of the students at this college who are Facebook users update their status at least two times a day? Test the relevant hypotheses using $\alpha = 0.05$.

b. Would it be reasonable to generalize the conclusion from the test in Part (a) to all college students in the United States? Explain why or why not.

10.95 A number of initiatives on the topic of legalized gambling have appeared on state ballots. A political candidate has decided to support legalization of casino gambling if they are convinced that more than two-thirds of adult Americans approve of casino gambling. Suppose that 1523 adults selected at random from households with telephones were asked whether they approved of casino gambling. The number in the sample who approved was 1035. Does the sample provide convincing evidence that more than two-thirds of adult Americans approve?

10.96 Duck hunting in populated areas faces opposition on the basis of safety and environmental issues. In a survey to assess public opinion regarding duck hunting on Morro Bay (located along the central coast of California), a random sample of 750 local residents included 560 who strongly opposed hunting on the bay. Does this sample provide sufficient evidence to conclude that the majority of local residents oppose hunting on Morro Bay? Test the relevant hypotheses using $\alpha = 0.01$.

10.97 Many people now screen incoming phone calls, which presents a problem for polling organizations. The article **"Who in the World is Still Answering Pollsters' Phone Calls?" (*The New York Times*, October 12, 2022)** indicated that in 2020, only about 20% of calls made by polling organizations resulted in an answered call. Suppose that you wanted to see if this percentage has changed since 2020. You monitor 1000 calls made by a polling organization to randomly generated phone numbers, and find that 149 resulted in an answered call. Would this be convincing evidence that the proportion of calls that result in an answered call is less than the 2020 proportion of 0.20? State and test appropriate hypotheses using a 0.05 significance level.

10.98 ● An automobile manufacturer who wishes to advertise that one of its models achieves 30 mpg (miles per gallon) decides to carry out a fuel efficiency test. Six nonprofessional drivers were selected, and each one drove a car from Phoenix to Los Angeles. The resulting fuel efficiencies (in miles per gallon) are:

27.2 29.3 31.2 28.4 30.3 29.6

Assuming that fuel efficiency is normally distributed under these circumstances, do the data contradict the claim that true average fuel efficiency is (at least) 30 mpg?

10.99 A student organization uses the proceeds from a particular soft-drink dispensing machine to finance its activities. The price per can had been \$0.75 for a long time, and the average daily revenue during that period was \$75.00. The price was recently increased to \$1.00 per can. A random sample of $n = 20$ days after the price increase yielded a sample mean daily revenue and sample standard deviation of \$70.00 and \$4.20, respectively.

Does this information suggest that the mean daily revenue has decreased from its value before the price increase? Test the appropriate hypotheses using $\alpha = 0.05$.

10.100 A hot tub manufacturer advertises that with its heating equipment, a temperature of 100°F can be achieved on average in 15 minutes or less. A random sample of 25 tubs is selected, and the time necessary to achieve a 100°F temperature is determined for each tub. The sample mean time and sample standard deviation are 17.5 minutes and 2.2 minutes, respectively. Does this information cast doubt on the company's claim? Carry out a test of hypotheses using significance level 0.05.

Copyright 2025 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s). Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

Technology Notes

z Test for Proportions

JMP

Summarized data

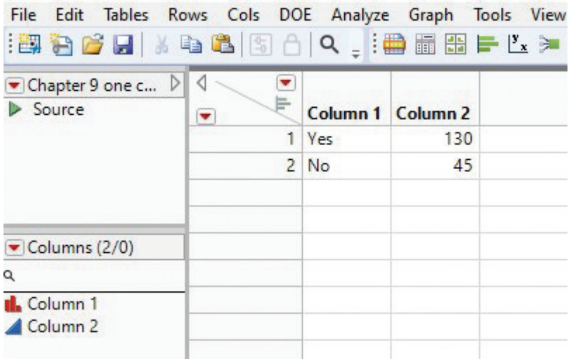
1. Enter the data into the JMP data table with categories in the first column and counts in the second column
2. Click **Analyze** and select **Distribution**
3. Click and drag the first column name from the box under **Select Columns** to the box next to **Y, Columns**
4. Click and drag the second column name from the box under **Select Columns** to the box next to **Freq**
5. Click **OK**
6. Click the red arrow next to the column name and click **Test Probabilities**
7. Under the **Test Probabilities** section that appears in the output, click in the box across from the categorical level representing a success, for example **Yes,** under **Hypoth Prob** and type the hypothesized value for p
8. Select the appropriate option for the alternative hypothesis
9. Click **Done**

Note: In the two-sided case, JMP uses the square of the z test statistic, called the Chi-Square test statistic. The two methods are mathematically identical.

Note: In the one-sided cases, JMP uses the exact binomial test rather than the z test.

Raw data

1. Enter the raw data into a column



	Column 1	Column 2
1	Yes	130
2	No	45

Source: JMP

2. Click **Analyze** and select **Distribution**
3. Click and drag the column name containing the data from the box under **Select Columns** to the box next to **Y, Columns**
4. Click **OK**
5. Click the red arrow next to the column name and click **Test Probabilities**
6. Under the **Test Probabilities** section that appears in the output, click in the box across from the categorical level representing a success, for example **Yes,** under **Hypoth Prob** and type the hypothesized value for p
7. Select the appropriate option for the alternative hypothesis
8. Click **Done**

Note: In the two-sided case, JMP uses the square of the z test statistic, called the Chi-Square test statistic. The two methods are mathematically identical.

Note: In the one-sided cases, JMP uses the exact binomial test rather than the z test.

Minitab

Summarized data

1. Click **Stat** then click **Basic Statistics** then click **1 Proportion…**
2. Select **Summarized data** from the dropdown menu
3. In the box next to **Number of Trials:** type the value for n, the sample size
4. In the box next to **Number of events:** type the value for the number of successes
5. Click the box next to **Perform hypothesis test** and input the appropriate hypothesized value in the box next to **Hypothesized proportion:**
6. Click **Options…**
7. Select the appropriate alternative hypothesis from the dropdown menu next to **Alternative hypothesis:**
8. Select **Normal approximation** from the dropdown menu next to **Method:**
9. Click **OK**
10. Click **OK**

Raw data

1. Input the raw data into a column
2. Click **Stat** then click **Basic Statistics** then click **1 Proportion…**
3. Select **One or more samples, each in a column** from the dropdown menu
4. Click on the box below the dropdown menu then double click the column name where the raw data is stored to copy the column name into the box
5. Click the box next to **Perform hypothesis test** and input the appropriate hypothesized value in the box next to **Hypothesized proportion:**
6. Click **Options…**
7. Select the appropriate alternative hypothesis from the dropdown menu next to **Alternative hypothesis:**
8. Select **Normal approximation** from the dropdown menu next to **Method:**
9. Click **OK**
10. Click **OK**

SPSS

SPSS does not have the functionality to automatically calculate a z test for a testing a single proportion.

Excel

Excel does not have the functionality to automatically calculate a z test for a testing a single population proportion. You may type in the formula by hand into a cell to have Excel calculate the value of the test statistic for you. Then use methods from Chapter 6 to find the P-value using the Normal Distribution.

TI-83/84

1. Press the **STAT** key
2. Highlight **TESTS**
3. Highlight **1-PropZTest…** and press **ENTER**

4. Next to **p0** type the hypothesized value for *p*
5. Next to **x** type the number of successes
6. Next to **n** type the sample size, *n*
7. Next to **prop,** highlight the appropriate alternative hypothesis
8. Highlight **Calculate** and press **ENTER**

TI-Nspire

1. Under **Scratchpad** select **A Calculate**
2. Press the **menu** key then select **6:Statistics** then select **7:Stat Tests** then **5:1-Prop z Test…** then press **enter**
3. In the box next to **p0** type the hypothesized value for *p*
4. In the box next to **Successes, x** type the number of successes
5. In the box next to **n** type the sample size, *n*
6. In the box next to **Alternate Hyp** choose the appropriate alternative hypothesis from the drop-down menu
7. Press **OK**

t Test for Population Mean, μ

JMP

1. Input the data into a column
2. Click **Analyze** and select **Distribution**
3. Click and drag the column name from the box under **Select Columns** to the box next to **Y, Response**
4. Click **OK**
5. Click the red arrow next to the column name and select **Test Mean**
6. In the box next to **Specify Hypothesized Mean**, type the hypothesized value of the mean, μ_0
7. Click **OK**

Note: The output provides results for all three possible alternative hypotheses.

Minitab

Summarized data

1. Click **Stat** then click **Basic Statistics** then click **1-sample t…**
2. Select **Summarized data** from the dropdown menu
3. In the box next to **Sample size:** type the value for *n*, the sample size
4. In the box next to **Mean:** type the value for the sample mean
5. In the box next to **Standard deviation:** type the value for the sample standard deviation
6. Click the box next to **Perform hypothesis** test and input the appropriate hypothesized value in the box next to **Hypothesized mean:**
7. Click **Options...**
8. Select the appropriate alternative hypothesis from the dropdown menu next to **Alternative hypothesis:**
9. Click **OK**
10. Click **OK**

Raw data

1. Input the raw data into a column
2. Click **Stat** then click **Basic Statistics** then click **1-sample t…**
3. Select **One or more samples, each in a column** from the dropdown menu
4. Click on the box below the dropdown menu then double click the column name where the raw data is stored to copy the column name into the box
5. Click the box next to **Perform hypothesis** test and input the appropriate hypothesized value in the box next to **Hypothesized mean:**
6. Click **Options...**
7. Select the appropriate alternative hypothesis from the dropdown menu next to **Alternative hypothesis:**
8. Select **Normal approximation** from the dropdown menu next to **Method:**
9. Click **OK**
10. Click **OK**

SPSS

1. Input the data into a column
2. Click **Analyze** then select **Compare Means** then select **One-Sample T Test…**
3. Highlight the column name for the variable
4. Click the arrow to move the variable to the **Test Variable(s):** box
5. In the box next to **Test Value:** input the hypothesized test value
6. Click **OK**

Note: This procedure produces a two-sided *P*-value.

Excel

Excel does not have the functionality to automatically produce a *t* test for a single population mean. However, you may type the formula into an empty cell manually to have Excel calculate the value of the test statistic for you. You can then use the steps below to find a *P*-value for the test statistic.

1. Select an empty cell
2. Click on **Formulas**
3. Click **Insert Function**
4. Select **Statistical** from the drop-down box for category
5. Select the appropriate *t*-distribution function (**Note: T.Dist** is the cumulative *t*-distribution $P(X \leq x)$; **T.Dist.2T** will return a two sided *P*-value; **T.Dist.RT** will return a one-sided right-tailed *P*-value)
6. Click in the box next to **X** and select the cell containing your test statistic or type it manually
7. Click in the box next to **Deg_freedom** and type the number of degrees of freedom $(n-1)$
8. Click **OK**

TI-83/84

Summarized data

1. Press **STAT**
2. Highlight **TESTS**
3. Highlight **T-Test…**
4. Highlight **Stats** and press **ENTER**
5. Next to $\boldsymbol{\mu_0}$ type the hypothesized value for the population mean
6. Next to $\bar{\mathbf{x}}$ input the value for the sample mean
7. Next to **Sx** input the value for the sample standard deviation
8. Next to **n** input the value for the sample size
9. Next to $\boldsymbol{\mu}$ highlight the appropriate alternative hypothesis and press **ENTER**
10. Highlight **Calculate** and press **ENTER**

Raw data

1. Enter the data into **L1** (In order to access lists press the **STAT** key, highlight the option called **Edit…** then press **ENTER**)
2. Press **STAT**
3. Highlight **TESTS**

Copyright 2025 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s). Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

4. Highlight **T-Test…**
5. Highlight **Data** and press **ENTER**
6. Next to μ_0 type the hypothesized value for the population mean
7. Next to μ highlight the appropriate alternative hypothesis and press **ENTER**
8. Highlight **Calculate** and press **ENTER**

TI-Nspire

Summarized data

1. Under **Scratchpad** select **A Calculate**
2. Press the **menu** key and select **6: Statistics** then **7: Stat Tests** then **2: t Test…** and press **enter**
3. From the drop-down menu select **Stats**
4. Press **OK**
5. Next to μ_0 type the hypothesized value for the population mean
6. Next to $\bar{x}$ input the value for the sample mean
7. Next to **Sx** input the value for the sample standard deviation
8. Next to **n** input the value for the sample size
9. Next to **Alternate Hyp** select the appropriate alternative hypothesis from the drop-down menu
10. Press **OK**

Raw data

1. Enter the data into a data list (In order to access data lists select the spreadsheet option and press **enter**)

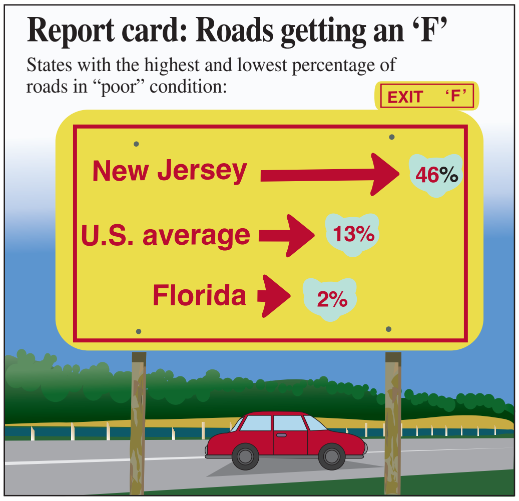
   Note: Be sure to title the list by selecting the top row of the column and typing a title.
2. Press the **menu** key and select **4: Statistics** then **4: Stat Tests** then **2: t Test…** and press **enter**
3. From the drop-down menu select **Data**
4. Press **OK**
5. Next to μ_0 input the hypothesized value for the population mean
6. Next to **List** select the list containing your data
7. Next to **Alternate Hyp** select the appropriate alternative hypothesis from the drop-down menu
8. Press **OK**

Cumulative Review Exercises CR10.1–CR10.16

● Data set available online

CR10.1 The *AARP Bulletin* (March 2010) included the following short news brief: "Older adults who did 1 hour of tai chi twice weekly cut their pain from knee osteoarthritis considerably in a 12-week study conducted at Tufts University School of Medicine." Suppose you were asked to design a study to investigate this claim. Describe an experiment that would allow comparison of the reduction in knee pain for those who did 1 hour of tai chi twice weekly to the reduction in knee pain for those who did not do tai chi. Include a discussion of how study participants would be selected, how pain reduction would be measured, and how participants would be assigned to experimental groups.

CR10.2 The following graphical display is similar to one that appeared in *USA TODAY*, (June 3, 2009). Write a few sentences critiquing this graphical display. Do you think it does a good job of creating a visual representation of the three percentages in the display?


Report card: Roads getting an 'F'
States with the highest and lowest percentage of roads in "poor" condition:
EXIT 'F'
New Jersey
46%
U.S. average
13%
Florida
2%

Sources: American Association of State Highway and Transportation Officials; USA TODAY. June 3, 2009.

CR10.3 ● The U.S. Bureau of Transportation Statistics included the accompanying data on the number of flights canceled in December 2021 for U.S. airlines.

Airline	Number of Flights	Rate per 1000 Flights
Alaska	1583	54.5
Allegiant	649	62.8
American	1763	11.3
Delta	2988	25.3
Frontier	212	16.1
Hawaiian	123	19.8
JetBlue	918	43.3
Southwest	1100	11.3
Spirit	871	48.4
United	3566	32.1

a. Construct a dotplot of the data on number of canceled flights. Are there any unusual observations that stand out in the dot plot? What airlines appear to be the worst in terms of number of canceled flights?

b. Construct a dotplot of the data on rate per 100,000 flights. Write a few sentences describing the interesting features of this plot.

c. If you wanted to compare airlines on the basis of flight cancellations, would you recommend using the data on number of flights canceled or on rate per 100,000 flights? Explain the reason for your choice.

CR10.4 ● The article "Wait Times on Rise to See Doctor" (*USA TODAY*, June 4, 2009) gave the accompanying data on average wait times in days to get an appointment with a medical specialist in 15 U.S. cities. Construct a boxplot of the average wait-time data. Are there any outliers in the data set?

Copyright 2025 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s). Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

City	Average Appointment Wait Time
Atlanta	11.2
Boston	49.6
Dallas	19.2
Denver	15.4
Detroit	12.0
Houston	23.4
Los Angeles	24.2
Miami	15.4
Minneapolis	19.8
New York	19.2
Philadelphia	27.0
Portland	14.4
San Diego	20.2
Seattle	14.2
Washington, D.C.	22.6

CR10.5 The report **"Online Security Survey, Google/Harris Poll" (February, 2019)** states that only 24% of Americans use a password manager, even though many people say they need a better way to track their passwords. For purposes of this exercise, assume that the 24% figure is correct for the population of adult Americans.

a. If a random sample of 20 adult Americans is selected, what is the probability that exactly 3 of them use a password manager?

b. What is the probability that more than 8 people in a random sample of 20 adult Americans use a password manager?

c. What is the mean and standard deviation of the variable x = number of people in a random sample of 100 adult Americans who use a password manager?

d. Find the approximate probability that the number of people who use a password manager in a random sample of 100 adult Americans is less than 20.

CR10.6 The article **"Should Canada Allow Direct-to-Consumer Advertising of Prescription Drugs?" (*Canadian Family Physician* [2009]: 130–131)** calls for the legalization of advertising of prescription drugs in Canada. Suppose you wanted to conduct a survey to estimate the proportion of Canadians who would support allowing this type of advertising. How large a random sample would be required to estimate this proportion to within 0.02 with 95% confidence?

CR10.7 **The National Association of Colleges and Employers** carries out a student survey each year. A summary of data from the 2021 survey included the following information:

- 22% of students graduating in 2021 intended to go on to graduate or professional school.
- Only 12% of those who graduated in 2021 attended a college with fewer than 5000 students
- Of those who attended a college with fewer than 5000 students, 65% were employed full time after graduation.

Consider the following events:

S = event that a randomly selected 2021 graduate attended a college with fewer than 5000 students
E = event that a randomly selected 2021 graduate was employed full time after graduation
G = event that a randomly selected 2021 graduate plans to attend graduate or professional school

Find the following probabilities.

a. $P(G)$
b. $P(S)$
c. $P(E|S)$
d. $P(E \cap S)$

CR10.8 It probably wouldn't surprise you to know that Valentine's Day means big business for florists, jewelry stores, and restaurants. But would it surprise you to know that it is also a big day for pet stores? In January 2023, the National Retail Federation conducted a survey of a sample of adult consumers that they believed was selected in a way that would produce a sample representative of the population of adults in the United States **("Valentine's Day Data Center," nrf.com)**. One of the questions in the survey asked if the respondent planned to spend money on a Valentine's Day gift for their pet this year.

a. The proportion who responded that they did plan to purchase a gift for their pet was 0.32. Suppose that the sample size for this survey was $n = 200$. Construct and interpret a 95% confidence interval for the proportion of all adult Americans who planned to purchase a Valentine's Day gift for their pet in 2023.

b. The actual sample size for the survey was much larger than 200. Would a 95% confidence interval computed using the actual sample size have been narrower or wider than the confidence interval computed in Part (a)?

c. Still assuming a sample size of $n = 200$, carry out a hypothesis test to determine if the data provides convincing evidence that the proportion who planned to buy a Valentine's Day gift for their pet in 2023 was greater than 0.25. Use a significance level of 0.05.

CR10.9 **"Many Americans Believe Fake News Is Sowing Confusion"** was the title of a press release issued by **Pew Research Center** after conducting a survey of 1002 Americans in December 2016. At the same time, the article reported that Americans are "generally confident in their ability to identify fake news," with 84% reporting that they are "very confident" or "somewhat confident" in their own ability to recognize made-up news. Suppose you want to determine if there is convincing evidence that a majority of all adult Americans are confident in their ability to recognize fake news.

a. What hypotheses should be tested in order to answer this question?

b. The P-value for this test is very small (< 0.00001). What conclusion would you reach if $\alpha = 0.01$?

Copyright 2025 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s). Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

c. A friend in your statistics class says that you do not have to worry about how the sample was selected when the *p*-value is so small, because the *p*-value provides extremely strong evidence. Do you agree or disagree?

CR10.10 **"Heinz Plays Catch-up After Under-Filling Ketchup Containers"** is the headline of an article that appeared on CNN.com **(November 30, 2000)**. The article stated that Heinz had agreed to put an extra 1% of ketchup into each ketchup container sold in California for a 1-year period. Suppose that you want to make sure that Heinz is in fact fulfilling its end of the agreement. You plan to take a sample of 20-oz bottles shipped to California, measure the amount of ketchup in each bottle, and then use the resulting data to estimate the mean amount of ketchup in 20-oz bottles. A small pilot study showed that the amount of ketchup in 20-oz bottles varied from 19.9 to 20.3 oz. How many bottles should be included in the sample if you want to estimate the true mean amount of ketchup to within 0.1 oz with 95% confidence?

CR10.11 Each year, the American Bar Association conducts a survey of civic literacy **(americanbar.org, 2023)**. In 2023, people in a nationally representative sample of 1000 adult Americans were asked if they thought that civility in today's society is worse compared to 10 years ago. The proportion who responded yes to this question was 0.85. Construct and interpret a 90% confidence interval for the proportion of adult Americans who believe that civility is worse compared to 10 years ago.

CR10.12 The survey of civic literacy referenced in the previous exercise also found that 530 of the 1000 people surveyed said that they believe that the general public was not very well informed about how the U.S. government works. Do these data provide convincing evidence that a majority of adult Americans share this belief? Test the relevant hypotheses using a significance level of 0.01.

CR10.13 The article titled **"13% of Americans Don't Use the Internet. Who Are They?"** describes a study conducted by the Pew Research Center **(pewresearch.org, September 7, 2016)**. Suppose that the title of this article is based on a representative sample of 600 adult Americans. Does this support the claim that the proportion of adult Americans who do not use the Internet is greater than 0.10 (10%)?

CR10.14 Suppose that a random sample of 50 cans of a particular brand of fruit juice is selected, and the amount of juice (in ounces) in each of the cans is determined. Let μ denote the mean amount of juice for the population of all cans of this brand. Suppose that this sample of 50 results in a 95% confidence interval for μ of (7.8, 9.4).

a. Would a 90% confidence interval have been narrower or wider than the given interval? Explain your answer.

b. Consider the following statement: There is a 95% chance that μ is between 7.8 and 9.4. Is this statement correct? Why or why not?

c. Consider the following statement: If the process of selecting a random sample of size 50 and then calculating the corresponding 95% confidence interval is repeated 100 times, exactly 95 of the resulting intervals will include μ. Is this statement correct? Why or why not?

CR10.15 The authors of the paper **"Serum Zinc Levels of Cord Blood: Relation to Birth Weight and Gestational Period" (*Journal of Trace Elements in Medicine and Biology* [2015]: 180–183)** carried out a study of zinc levels of low-birth-weight babies and normal-birth-weight babies. For a sample of 50 low-birth-weight babies, the sample mean zinc level was 17.00 and the standard error $\left(\frac{s}{\sqrt{n}}\right)$ was 0.43. For a sample of 73 normal-birth-weight babies, the sample mean zinc level was 18.16 and the standard error was 0.32. Explain why the two standard errors are not the same.

CR10.16 The paper **"Relationship Between Food Waste, Diet Quality, and Environmental Sustainability" (*PLOS ONE*, April 18, 2018)** summarizes data from a study of food waste in the United States. Data from a representative sample of approximately 5000 adult Americans resulted in a mean food waste per day of 422 grams. For purposes of this exercise, assume that the sample size was 500 and that the sample standard deviation was 530 grams (this is consistent with summary values given in the paper). Construct and interpret a 98% confidence interval for the mean food waste weight for adult Americans.

Copyright 2025 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s). Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

Chapter

11

Comparing Two Populations or Treatments

Peter Muller/Cultura/Getty Images

Many studies are carried out to answer questions about the differences between two populations or two treatments. For example, the authors of the paper **"Measuring Actual Learning Versus Feeling of Learning in Response to Being Actively Engaged in the Classroom" (*Proceedings of the National Academy of Sciences*, September 4, 2019)** were interested in determining if students learn more when they are actively engaged in the classroom than they do from a passive lecture and whether students think they learn more in active classroom settings. The participants in this study were students enrolled in an introductory physics course. The students were randomly assigned to one of two groups during the week that statics was to be covered. Students in both groups received the same handouts with sample problems targeting key concepts. In the passive learning group, students listened to a lecture in which solutions to the sample problems were presented and students took notes. In the active learning group, students worked together to solve the sample problems, with the instructor offering assistance as needed. At the end of the week, students in both groups took a test on the material covered and they also provided a response to a question about how strongly they felt that they had learned a great deal during the week.

Using methods that you will learn in this chapter, the researchers were able to conclude that the average test score was greater for the active learning treatment. But they also found that students' "feeling of learning" was not as great for the active learning treatment. This led the researchers to conclude that students actually learn more in an active learning setting but feel like they learn less!

In this chapter, you will see hypothesis tests and confidence intervals that can be used to compare two populations or treatments.

Learning Objectives

After completing this chapter, you will be able to:

- Explain the difference between paired and independent samples.
- Distinguish between paired and independent samples.
- Carry out a test for a difference in means using independent samples or paired samples and interpret the results in context.

- Carry out a test for a difference in proportions using independent samples and interpret the results in context.
- Construct a confidence interval estimate of a difference in means using independent samples or paired samples and interpret the interval in context.
- Construct a confidence interval estimate of a difference in proportions using independent samples and interpret the interval in context.
- Calculate and interpret a bootstrap confidence interval for a difference in proportions. (Optional)
- Carry out a randomization test for a difference in proportions. (Optional)
- Calculate and interpret a bootstrap confidence interval for a difference in means. (Optional)
- Carry out a randomization test for a difference in means. (Optional)

Section 11.1

Inferences Concerning the Difference Between Two Population or Treatment Means Using Independent Samples

In this section, you will see how sample data is used to compare two population means or two treatment means. For example, the report **"Common Sense Census: Media Use by Kids Age Zero to Eight" (Common Sense Media, 2020)** summarizes data from surveys of a representative sample of parents of children of age 2 to 4 and a representative sample of parents of children of age 5 to 8. Data from these two samples could be used to learn about whether the mean time spent using mobile devices is greater for children in the older age group than for children in the younger age group. In this section, you will see how sample data can be used to answer questions like this that involve comparing two population means.

In other situations, an experiment might be carried out to compare two different treatments or to compare the effect of a treatment with the effect of no treatment. For example, an agricultural researcher might want to compare weight gains for animals placed on two different diets (each diet is a treatment). An educational researcher might want to compare online instruction to traditional classroom instruction by studying the difference in mean scores on a common final exam (each type of instruction is a treatment).

In previous chapters, the symbol μ was used to denote the mean of a single population under study. When comparing two populations or treatments, you will use notation that distinguishes between the characteristics of the first and those of the second. This is accomplished by using subscripts, as shown in the accompanying box.

Notation

	Mean	Variance	Standard Deviation
Population or Treatment 1	μ_1	σ_1^2	σ_1
Population or Treatment 2	μ_2	σ_2^2	σ_2

	Sample Size	Mean	Variance	Standard Deviation
Sample from Population or Treatment 1	n_1	$\bar{x}_1$	s_1^2	s_1
Sample from Population or Treatment 2	n_2	$\bar{x}_2$	s_2^2	s_2

Copyright 2025 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s). Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

A comparison of means focuses on the difference, $\mu_1 - \mu_2$. When $\mu_1 - \mu_2 = 0$, the two population or treatment means are equal. That is,

$$\mu_1 - \mu_2 = 0 \text{ is equivalent to } \mu_1 = \mu_2$$

Similarly,

$$\mu_1 - \mu_2 > 0 \text{ is equivalent to } \mu_1 > \mu_2$$

and

$$\mu_1 - \mu_2 < 0 \text{ is equivalent to } \mu_1 < \mu_2$$

Before developing methods for drawing conclusions about $\mu_1 - \mu_2$, it is important to consider how the two samples, one from each population or treatment, are selected. Two samples are said to be **independent** samples if the selection of the individuals or objects that make up one sample does not influence the selection of individuals or objects in the other sample.

When observations from the first sample are linked in some meaningful way with observations in the second sample, the samples are said to be **paired**. For example, to study the effectiveness of a keyboarding course, the typing speed of subjects could be measured before they take the course and again after they complete the course. This results in two related samples—one from the population of individuals who have not taken this particular course (the "before" measurements) and one from the population of individuals who have taken the course (the "after" measurements). These samples are paired. The two samples are not independently chosen, because the selection of individuals from the first (before) population completely determines which individuals make up the sample from the second (after) population. In Section 11.1, procedures based on independent samples are considered. Methods for analyzing data resulting from paired samples are presented in Section 11.2.

Because $\bar{x}_1$ provides an estimate of μ_1 and $\bar{x}_2$ gives an estimate of μ_2, it is natural to use $\bar{x}_1 - \bar{x}_2$ as a point estimate of $\mu_1 - \mu_2$. The value of $\bar{x}_1$ varies from sample to sample, as does the value of $\bar{x}_2$. Since the difference $\bar{x}_1 - \bar{x}_2$ is calculated from sample values, it is a statistic that has sample-to-sample variability. This variability is described by its sampling distribution.

Properties of the Sampling Distribution of $\bar{x}_1 - \bar{x}_2$

For independent random samples:

1. $$\mu_{\bar{x}_1 - \bar{x}_2} = \begin{pmatrix} \text{mean value} \\ \text{of } \bar{x}_1 - \bar{x}_2 \end{pmatrix} = \mu_{\bar{x}_1} - \mu_{\bar{x}_2} = \mu_1 - \mu_2$$

   This means that the sampling distribution of $\bar{x}_1 - \bar{x}_2$ is always centered at the value of $\mu_1 - \mu_2$, so $\bar{x}_1 - \bar{x}_2$ is an unbiased statistic for estimating $\mu_1 - \mu_2$.

2. $$\sigma^2_{\bar{x}_1 - \bar{x}_2} = \begin{pmatrix} \text{variance of} \\ \bar{x}_1 - \bar{x}_2 \end{pmatrix} = \sigma^2_{\bar{x}_1} + \sigma^2_{\bar{x}_2} = \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}$$

   and

   $$\sigma_{\bar{x}_1 - \bar{x}_2} = \begin{pmatrix} \text{standard deviation} \\ \text{of } \bar{x}_1 - \bar{x}_2 \end{pmatrix} = \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$$

3. If n_1 and n_2 are both large or the population distributions are (at least approximately) normal, $\bar{x}_1$ and $\bar{x}_2$ each have (at least approximately) a normal distribution. This implies that the sampling distribution of $\bar{x}_1 - \bar{x}_2$ is also normal or approximately normal.

Copyright 2025 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s). Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

Properties 1 and 2 follow from the following general results:

1. The mean value of a difference in means is the difference of the two individual mean values.
2. The variance of a difference of *independent* quantities is the *sum* of the two individual variances.

When the sample sizes are large or when the population distributions are approximately normal, the properties of the sampling distribution of $\bar{x}_1 - \bar{x}_2$ imply that $\bar{x}_1 - \bar{x}_2$ can be standardized to obtain a variable with a sampling distribution that is approximately the standard normal (z) distribution. This leads to the following result.

When two random samples are independently selected and when n_1 and n_2 are both large or the population distributions are normal, the statistic

$$z = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{\sqrt{\dfrac{\sigma_1^2}{n_1} + \dfrac{\sigma_2^2}{n_2}}}$$

has a standard normal (z) distribution.

What Do You Think?

In most situations, you would not be able to calculate the value of the z statistic given in the box above. Why is this the case?

Although it is possible to base a test procedure and confidence interval on this result, the values of σ_1^2 and σ_2^2 are usually not known. When σ_1^2 and σ_2^2 are not known, you can estimate them using the corresponding sample variances, s_1^2 and s_2^2. This leads to the result given in the accompanying box.

When two random samples are independently selected and when n_1 and n_2 are both large or when the population distributions are normal, the standardized variable

$$t = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{\sqrt{\dfrac{s_1^2}{n_1} + \dfrac{s_2^2}{n_2}}}$$

has approximately a t distribution with

$$\text{df} = \frac{(V_1 + V_2)^2}{\dfrac{V_1^2}{n_1 - 1} + \dfrac{V_2^2}{n_2 - 1}} \quad \text{where} \quad V_1 = \frac{s_1^2}{n_1} \quad \text{and} \quad V_2 = \frac{s_2^2}{n_2}$$

The calculated value of df should be truncated (rounded down) to obtain an integer value of df.

If one or both sample sizes are small, you can use normal probability plots or boxplots to evaluate whether it is reasonable to consider the population distributions to be approximately normal.

Hypothesis Tests

In a test designed to compare two population means, the null hypothesis is of the form

$$H_0\text{: } \mu_1 - \mu_2 = \text{hypothesized value}$$

Often the hypothesized value is 0, indicating that there is no difference between the population means. The alternative hypothesis involves the same hypothesized value but uses one of three inequalities (less than, greater than, or not equal to), depending on the research question of interest. As an example, suppose μ_1 and μ_2 denote the average fuel efficiencies (in miles per gallon, mpg) for cars with 4-cylinder engines and cars with 6-cylinder engines, respectively. The hypotheses under consideration might be

$$H_0\text{: } \mu_1 - \mu_2 = 5 \quad \text{versus} \quad H_a\text{: } \mu_1 - \mu_2 > 5$$

The null hypothesis is equivalent to the claim that the mean fuel efficiency for the 4-cylinder engine cars exceeds the mean fuel efficiency for the 6-cylinder engine cars by 5 mpg. The alternative hypothesis states that the difference between the mean fuel efficiencies is more than 5 mpg.

A test statistic is obtained by replacing $\mu_1 - \mu_2$ in the standardized t variable (given in the previous box) with the hypothesized value that appears in H_0. For example, the t statistic for testing H_0: $\mu_1 - \mu_2 = 5$ is

$$t = \frac{(\bar{x}_1 - \bar{x}_2) - 5}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

When the sample sizes are large or when the population distributions are normal, the sampling distribution of the test statistic is approximately a t distribution when H_0 is true.

Two-Sample *t* Test for Comparing Two Population Means

Null Hypothesis: H_0: $\mu_1 - \mu_2 = \text{hypothesized value}$

Test Statistic: $t = \dfrac{(\bar{x}_1 - \bar{x}_2) - \text{hypothesized value}}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$

The appropriate df for the two-sample t test is

$$\text{df} = \frac{(V_1 + V_2)^2}{\frac{V_1^2}{n_1 - 1} + \frac{V_2^2}{n_2 - 1}} \quad \text{where} \quad V_1 = \frac{s_1^2}{n_1} \quad \text{and} \quad V_2 = \frac{s_2^2}{n_2}$$

The calculated number of degrees of freedom should be truncated (rounded down) to an integer.

Alternative Hypothesis:	***P*-Value:**
H_a: $\mu_1 - \mu_2 > \text{hypothesized value}$	Area under appropriate t curve to the right of t
H_a: $\mu_1 - \mu_2 < \text{hypothesized value}$	Area under appropriate t curve to the left of t
H_a: $\mu_1 - \mu_2 \neq \text{hypothesized value}$	2(area to the right of t) if t is positive or 2(area to the left of t) if t is negative

Assumptions:
1. The two samples are independently selected random samples from the populations of interest.
2. The sample sizes are large (generally 30 or larger) or the population distributions are (at least approximately) normal.

The *P*-value for the test is obtained by first calculating the appropriate number of degrees of freedom and then using Appendix Table 4, a graphing calculator, or a statistical software package.

Example 11.1 Lead Exposure of Bald Eagles and Golden Eagles

In a study to investigate the extent of exposure to lead in predatory birds, the authors of the paper **"Demographic Implications of Lead Poisoning for Eagles Across North America" (*Science* [2022]: 779–782)** measured blood lead concentration (in mg/dL) for birds in a representative sample of 145 adult bald eagles and in an independently selected representative sample of 116 adult golden eagles.

Means and standard deviations for these samples are given in the accompanying table.

Population	Population Mean	Sample Size	Sample Mean	Sample Standard Deviation
Bald eagles	μ_1 = mean blood lead concentration for bald eagles	$n_1 = 145$	$\bar{x}_1 = 92.78$	$s_1 = 264.79$
Golden eagles	μ_2 = mean blood lead concentration for golden eagles	$n_2 = 116$	$\bar{x}_2 = 22.78$	$s_2 = 40.84$

What Do You Think?

Notice that for both samples, the standard deviation is greater than the mean. The value of blood lead concentration can't be negative. What does this imply about the distribution of values in each of the two samples?

Do these data provide convincing evidence that the mean blood lead concentration is different for bald eagles and golden eagles? You can answer this question by testing the relevant hypotheses using a 0.05 level of significance.

Understand the context ➤

1. μ_1 = population mean blood lead concentration for bald eagles
   μ_2 = population mean blood lead concentration for golden eagles
   $\mu_1 - \mu_2$ = difference in mean blood lead concentration
2. H_0: $\mu_1 - \mu_2 = 0$ (no difference in mean blood lead concentration)
3. H_a: $\mu_1 - \mu_2 \neq 0$ (mean blood lead concentration is different for bald eagles and golden eagles)
4. $\alpha = 0.05$

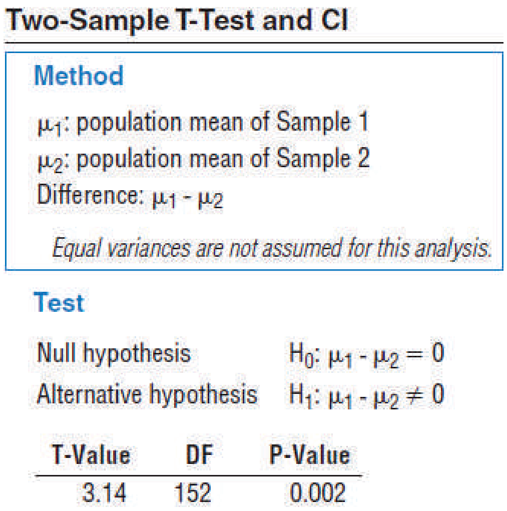
Formulate a plan ➤

5. Test statistic: $t = \dfrac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{\sqrt{\dfrac{s_1^2}{n_1} + \dfrac{s_2^2}{n_2}}} = \dfrac{(\bar{x}_1 - \bar{x}_2) - 0}{\sqrt{\dfrac{s_1^2}{n_1} + \dfrac{s_2^2}{n_2}}}$
6. Assumptions: There are two assumptions that need to be considered in order to use the two-sample *t* test. It is not reasonable to think that the population distributions are approximately normal. Blood lead concentration can't be negative and the sample standard deviations are both much larger than the sample means. This would not be possible if the distributions were approximately normal and indicates that the distributions are skewed with a long upper tail. However, the large-sample condition is met because the sample sizes are both large ($n_1 = 145 \geq 30$ and $n_2 = 116 \geq 30$). From the study description, you know that the samples were independently selected. You also know that the samples were selected to be representative of the two populations of interest. The assumptions are reasonable, so use of the two-sample *t* test is appropriate.

Do the work ➤

7. Calculation: A statistics software package or graphing calculator can be used to calculate the value of the test statistic. For example, Minitab output is shown here.

Copyright 2025 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s). Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.



```
Two-Sample T-Test and CI

Method

μ1: population mean of Sample 1
μ2: population mean of Sample 2
Difference: μ1 - μ2

Equal variances are not assumed for this analysis.

Test

Null hypothesis          H0: μ1 - μ2 = 0
Alternative hypothesis   H1: μ1 - μ2 ≠ 0

T-Value   DF    P-Value
   3.14  152      0.002
```

The value of the test statistic is $t = 3.14$.

8. P-value: From the Minitab output, P-value $= 0.002$

   If you don't have access to technology and must calculate by hand, those calculations are illustrated below.

   Test statistic:

$$t = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} = \frac{(92.78 - 22.78) - 0}{\sqrt{\frac{(264.79)^2}{145} + \frac{(40.84)^2}{116}}}$$

$$= \frac{70.00}{22.31}$$

$$= 3.14$$

   Degrees of freedom:

$$V_1 = \frac{s_1^2}{n_1} = 483.54 \qquad V_2 = \frac{s_2^2}{n_2} = 14.38$$

$$\text{df} = \frac{(V_1 + V_2)^2}{\frac{V_1^2}{n_1 - 1} + \frac{V_2^2}{n_2 - 1}}$$

$$= \frac{(483.54 + 14.38)^2}{\frac{(483.54)^2}{144} + \frac{(14.38)^2}{115}}$$

$$= \frac{247{,}924.33}{1{,}625.49}$$

$$= 152.52$$

   Truncate to df $= 152$

   This is a two-tailed test (the inequality in H_a is $\neq$), so the P-value is two times the area to the right of 3.14 under the t curve with df $= 152$. Using the ∞ df row of Appendix Table 4, the area to the right of 3.14 is approximately 0.001.

   Associated P-value:

$$P\text{-value} = 2(0.001) = 0.002$$

Interpret the results ➤

9. Conclusion: Because the P-value is less than the selected significance level, you reject the null hypothesis. Based on the sample data, there is convincing evidence that the mean blood lead concentration is not the same for bald eagles and golden eagles.

You can conclude that the sample mean blood lead concentration for bald eagles is different enough from the sample mean for golden eagles that you don't think that this could have occurred just by chance when there is no difference in the population means.

Example 11.2 Rental Frogs?

● A frog jumping competition described in a short story written by Mark Twain inspired the real annual Calaveras County Frog Jumping Jubilee. In this competition, people enter bullfrogs into a contest to see which frog can jump the farthest. Some serious competitors have frogs that they have trained, and these contestants are known as "professional frogs." Amateurs also compete with frogs that they can rent from the contest organizers, and these contestants are known as "rental frogs." The authors of the paper **"Chasing Maximal Performance: A Cautionary Tale from the Celebrated Jumping Frogs of Calaveras County" (*Journal of Experimental Biology* [2013]: 3947–3953)** wanted to compare the performance of rental frogs and professional frogs. The authors of the paper used a two-sample *t* test to compare the mean jump distance (in meters) for the two groups.

Data consistent with summary quantities given in the paper are shown in the accompanying table (the actual sample sizes in the study were much larger).

	Jump Distance (in meters)														
Rental Frog	0.9	0.9	1.3	0.8	1.1	1.2	0.8	0.9	1.4	1.1	1.1	1.0	0.9	1.2	1.2
Professional Frog	1.0	1.5	2.2	1.2	1.4	1.3	1.5	1.9	1.7	1.6	0.7	1.2	1.3	1.0	0.5

The authors of the paper hypothesized that rental frogs would not perform as well as professional frogs. Assuming that the frogs in the two samples are representative of rental frogs and professional frogs in general, the sample data can be used to determine if there is convincing evidence that the mean jumping distance for rental frogs is less than the mean distance for professional frogs.

Understand the context ➤

1. μ_1 = population mean jump distance for rental frogs
μ_2 = population mean jump distance for professional frogs
$\mu_1 - \mu_2$ = difference in mean jump distance

2. H_0: $\mu_1 - \mu_2 = 0$

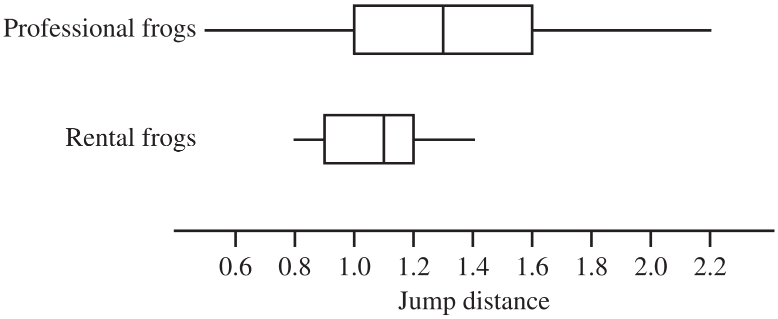
3. H_a: $\mu_1 - \mu_2 < 0$

Formulate a plan ➤

4. Significance level: $\alpha = 0.05$

5. Test statistic: $$t = \frac{(\bar{x}_1 - \bar{x}_2) - \text{hypothesized value}}{\sqrt{\dfrac{s_1^2}{n_1} + \dfrac{s_2^2}{n_2}}} = \frac{(\bar{x}_1 - \bar{x}_2) - 0}{\sqrt{\dfrac{s_1^2}{n_1} + \dfrac{s_2^2}{n_2}}}$$

6. Assumptions: The samples were independently selected, and you can assume that the samples are representative of the populations of interest. Because both of the sample sizes are small, it is also necessary to assume that the jump distance distribution is approximately normal for each of the two populations. Boxplots constructed using the sample data are shown here:



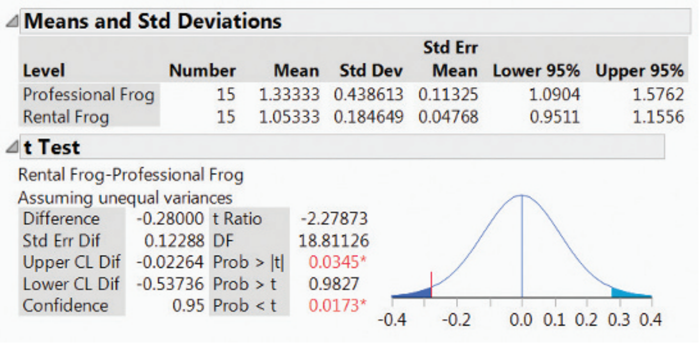
● **Data set available online**

Copyright 2025 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s). Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

Because the boxplots are reasonably symmetric and because there are no outliers, it is reasonable to assume that the two population distributions are approximately normal. With all of the assumptions met, it is appropriate to proceed with the two-sample *t*-test.

Do the work ➤

7. Calculation: JMP was used to do the calculations, resulting in the following output. From the JMP output, the value of the test statistic is -2.28 (rounded from -2.27873 in the output labeled t Ratio).



Means and Std Deviations

Level	Number	Mean	Std Dev	Std Err Mean	Lower 95%	Upper 95%
Professional Frog	15	1.33333	0.438613	0.11325	1.0904	1.5762
Rental Frog	15	1.05333	0.184649	0.04768	0.9511	1.1556

t Test

Rental Frog-Professional Frog

Assuming unequal variances

Difference	-0.28000	t Ratio	-2.27873
Std Err Dif	0.12288	DF	18.81126
Upper CL Dif	-0.02264	Prob > |t|	0.0345*
Lower CL Dif	-0.53736	Prob > t	0.9827
Confidence	0.95	Prob < t	0.0173*

8. From the JMP output given in Step 7, df = 18 (rounded down from 18.81126 in the output labeled DF). The associated *P*-value is 0.0173 (from the Prob $< t$ entry in the output because this is a lower tail test).

Interpret the results ➤

9. Conclusion: Because the *P*-value of 0.0173 is less than 0.05, you reject H_0. The sample data provide convincing evidence that the mean jump distance for rental frogs is less than the mean jump distance for professional frogs.

Suppose the value of the test statistic in Step 7 had been -1.28 rather than -2.28. Then the *P*-value would have been 0.105 (the area to the left of -1.3 under the *t* curve with 18 df) and the decision would have been to not reject the null hypothesis. You then would have concluded that there was not convincing evidence that the mean jump distance for rental frogs was less than the mean for professional frogs. Notice that when you fail to reject the null hypothesis of no difference between the population means, you are not saying that there is convincing evidence that the means are equal—you can only say that you were not convinced that they were different.

Comparing Treatments

When an experiment is carried out to compare two treatments (or to compare a single treatment with a control), you are interested in the effect of the treatments on some response variable. The treatments are "applied" to individuals (as in an experiment to compare two different medications for decreasing blood pressure) or objects (as in an experiment to compare two different baking temperatures on the density of bread). The value of some response variable (for example, blood pressure or density) is recorded. Based on the resulting data, you can determine whether there is a significant difference in the mean response for the two treatments.

In many actual experimental situations, the individuals or objects to which the treatments will be applied are not selected at random from some larger population. A consequence of this is that it is not possible to generalize the results of the experiment to some larger population. However, if individuals or objects are *randomly assigned* to treatments (or treatments are randomly assigned to individuals or objects), it is possible to test hypotheses about treatment differences.

It is common practice to use the two-sample *t* test statistic previously described if the experiment employs random assignment and if either the sample sizes are large or it is reasonable to think that the treatment response distributions (the distributions of response values that would result if the treatments were applied to a *very* large number of individuals or objects) are approximately normal.

Two-Sample *t* Test for Comparing Two Treatment Means

When

1. individuals or objects are randomly assigned to treatments (or treatments are randomly assigned to individuals or objects), and
2. the sample sizes are large (generally 30 or larger) or the treatment response distributions are approximately normal,

the two-sample *t* test can be used to test

$$H_0: \mu_1 - \mu_2 = \textit{hypothesized value}$$

where μ_1 and μ_2 represent the mean response for treatments 1 and 2, respectively.

In this case, these two conditions replace the assumptions previously stated for comparing two population means. Whether the assumption of normality of the treatment response distributions is reasonable can be assessed by constructing normal probability plots or boxplots of the response values in each sample.

When the two-sample *t* test is used to compare two treatments when the individuals or objects used in the experiment are not randomly selected from some population, it is only an approximate test (the reported *P*-values are only approximate). However, this is still the most common way to analyze such data.

Andrey_Popov/Shutterstock.com

Example 11.3 Fitness Trackers and Weight Loss

The article **"Activity Trackers May Undermine Weight Loss Efforts" (*The New York Times*, September 20, 2016)** describes a study published in the ***Journal of the American Medical Association* ("The Effect of Wearable Technology Combined with a Lifestyle Intervention on Long-Term Weight Loss." [2016]: 1161–1171)**. In this study, subjects followed a low-calorie diet and exercise program for 6 months. After 6 months, the subjects were randomly assigned to one of two groups. The people in one group were provided with a website they could use to self-monitor diet and physical activity. The people in the second group were provided with a wearable fitness tracker with an accompanying web interface to monitor diet and physical activity.

The researchers were interested in learning if the mean weight loss (in kilograms) at the end of 2 years was different for the two treatments (self-monitoring and fitness tracker monitoring). Data from this experiment are summarized in the accompanying table.

Group	Sample Size	Mean Weight Loss	Standard Deviation
Self-Monitoring	170	5.9 kg	6.8 kg
Fitness Tracker Monitoring	181	3.5 kg	6.3 kg

Do the data from this experiment provide evidence that the mean weight loss differs for the two treatments? You can test the relevant hypotheses using a significance level of 0.01.

Understand the context ➤

1. Let μ_1 denote the mean weight loss for the self-monitor treatment and let μ_2 denote the mean weight loss for the fitness tracker treatment. Then $\mu_1 - \mu_2$ is the difference in treatment means.
2. H_0: $\mu_1 - \mu_2 = 0$
3. H_a: $\mu_1 - \mu_2 \neq 0$
4. Significance level: $\alpha = 0.01$

Formulate a plan ➤

5. Test statistic: $t = \dfrac{(\bar{x}_1 - \bar{x}_2) - \text{hypothesized value}}{\sqrt{\dfrac{s_1^2}{n_1} + \dfrac{s_2^2}{n_2}}} = \dfrac{(\bar{x}_1 - \bar{x}_2) - 0}{\sqrt{\dfrac{s_1^2}{n_1} + \dfrac{s_2^2}{n_2}}}$
6. Assumptions: Subjects were randomly assigned to the treatment groups, and both sample sizes are large, so use of the two-sample *t* test is reasonable.

Copyright 2025 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s). Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

Do the work ➤

7. Calculation: $$t = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} = \frac{(5.9 - 3.5) - 0}{\sqrt{\frac{(6.8)^2}{170} + \frac{(6.3)^2}{181}}} = \frac{2.4}{0.701} = 3.42$$

8. P-value: We first calculate the df for the two-sample t test:

$$V_1 = \frac{s_1^2}{n_1} = \frac{(6.8)^2}{170} = 0.272 \qquad V_2 = \frac{s_2^2}{n_2} = \frac{(6.3)^2}{181} = 0.219$$

$$\text{df} = \frac{(V_1 + V_2)^2}{\frac{V_1^2}{n_1 - 1} + \frac{V_2^2}{n_2 - 1}} = \frac{(0.272 + 0.219)^2}{\frac{(0.272)^2}{169} + \frac{(0.219)^2}{180}} = \frac{0.241}{0.0007} = 344.286$$

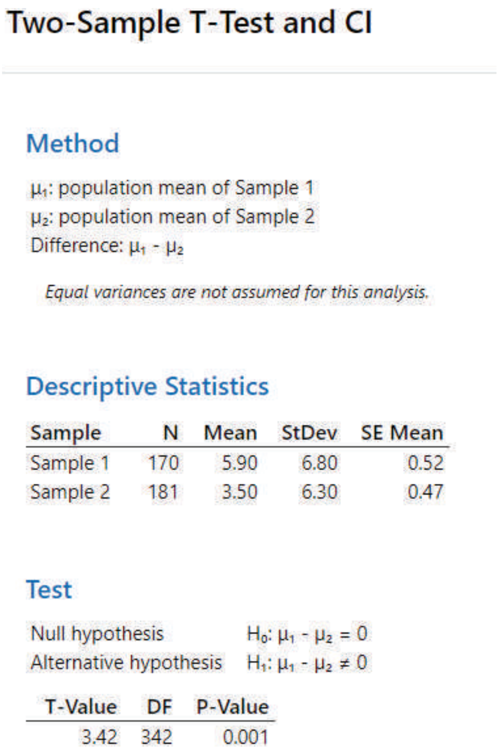
Truncating the calculated degrees of freedom results in df = 344.

This is a two-tailed test, so the P-value is two times the area under the t curve with df = 344 and to the right of 3.42. Since 3.42 is so far out in the upper tail of this t curve, P-value ≈ 0.

Interpret the results ➤

9. Conclusion: Because the P-value is less than the selected significance level (0.01), H_0 is rejected. There is convincing evidence that the mean weight loss is not the same for the self-monitoring treatment and the fitness tracker monitoring treatment. Notice that the mean weight loss for the group that used the fitness tracker was less than the mean for the group that self monitored. This is the basis for the headline from *The New York Times* article that said that fitness trackers may undermine weight loss efforts.

A statistical software package or a graphing calculator could also have been used to complete the Calculate step. For example, Minitab output is shown here. From the Minitab output, the value of the test statistic is $t = 3.42$. The reported degrees of freedom is 342 and the P-value is reported as 0.001. The small difference in the value of the degrees of freedom from the output and from the hand calculations is due to differences in rounding.



```
Two-Sample T-Test and CI

Method

μ₁: population mean of Sample 1
μ₂: population mean of Sample 2
Difference: μ₁ - μ₂

  Equal variances are not assumed for this analysis.

Descriptive Statistics

Sample      N  Mean  StDev  SE Mean
Sample 1  170  5.90   6.80     0.52
Sample 2  181  3.50   6.30     0.47

Test

Null hypothesis         H₀: μ₁ - μ₂ = 0
Alternative hypothesis  H₁: μ₁ - μ₂ ≠ 0

 T-Value   DF  P-Value
    3.42  342    0.001
```

Copyright 2025 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s). Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

You have probably noticed that evaluating the formula for number of degrees of freedom for the two-sample t test involves quite a bit of arithmetic. An alternative approach is to calculate a conservative estimate of the P-value—one that is close to but larger than the actual P-value. If H_0 is rejected using this conservative estimate, then it will also be rejected if the actual P-value is used.

> A conservative estimate of the P-value for the two-sample t test can be found by using the t curve with the number of degrees of freedom equal to the smaller of $(n_1 - 1)$ and $(n_2 - 1)$.

The Pooled *t* Test

The two-sample t test just described is appropriate when it is reasonable to assume that the population distributions are approximately normal. If it is also known that the variances of the two populations are equal ($\sigma_1^2 = \sigma_2^2$), an alternative test known as the *pooled t test* can be used. This test combines information from both samples to obtain a "pooled" estimate of the common variance and then uses this pooled estimate of the variance in place of s_1^2 and s_2^2 in the t test statistic.

The pooled t test was widely used in the past, but it has fallen into some disfavor because it is quite sensitive to departures from the assumption of equal population variances. If the population variances are equal, the pooled t test has a slightly better chance of detecting departures from H_0 than does the two-sample t test of this section. However, P-values based on the pooled t test can be seriously in error if the population variances are not equal, so, in general, the two-sample t test is a better choice than the pooled t test.

Comparisons and Causation

If the assignment of treatments to the individuals or objects used in a comparison of treatments is not made by the investigators, the study is observational. As an example, the article **"Lead and Cadmium Absorption Among Children near a Nonferrous Metal Plant" (*Environmental Research* [1978]: 290–308)** reported data on blood lead concentrations for two different samples of children. The first sample was drawn from a population residing within 1 km of a lead smelter. Those in the second sample were selected from a rural area much farther from the smelter.

It was the parents of the children, rather than the investigators, who determined whether the children would be in the close-to-smelter group or the far-from-smelter group. As a second example, a letter in the ***Journal of the American Medical Association* (May 19, 1978)** reported on a comparison of doctors' longevity after medical school graduation for those with an academic affiliation and those in private practice. (The letter writer's stated objective was to see whether "publish or perish" really meant "publish *and* perish.") Here again, an investigator did not start out with a group of doctors, assigning some to academic and others to nonacademic careers. The doctors themselves selected their groups.

The difficulty with drawing conclusions based on an observational study is that a statistically significant difference may be due to some underlying factors that have not been controlled rather than to conditions that define the groups. Does the type of medical practice itself have an effect on longevity, or is the observed difference in mean lifetime caused by other factors, which themselves led graduates to choose academic or nonacademic careers? Similarly, is the observed difference in blood lead concentration levels due to proximity to the smelter? Perhaps other physical and socioeconomic factors are related both to choice of living area and to blood lead concentration.

In general, rejection of H_0: $\mu_1 - \mu_2 = 0$ in favor of H_a: $\mu_1 - \mu_2 > 0$ suggests that, on average, higher values of the variable are *associated* with individuals in the first population or receiving the first treatment than with those in the second population or receiving the second treatment. But association does not imply causation.

Strong statistical evidence for a causal relationship can be built up over time through many different comparative studies that point to the same conclusions (as in the many investigations linking smoking to lung cancer). A randomized controlled experiment, in which investigators assign subjects at random to the treatments or conditions being compared, is particularly effective in suggesting causality. With random assignment, the investigator and other interested parties can be more easily convinced that an observed difference is caused by the difference in treatments or conditions.

A Confidence Interval

A confidence interval for $\mu_1 - \mu_2$ is easily obtained from the basic t variable developed in this section. Both the derivation of and the formula for the interval are similar to those of the one-sample t interval introduced in Chapter 9.

Two-Sample *t* Confidence Interval for the Difference in Two Population or Treatment Means

The general formula for a confidence interval for $\mu_1 - \mu_2$ when

1. the two samples are independently chosen random samples
2. the sample sizes are both large (generally $n_1 \geq 30$ and $n_2 \geq 30$
   or
   the population distributions are approximately normal

is

$$(\bar{x}_1 - \bar{x}_2) \pm (t \text{ critical value})\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$$

The t critical value is based on

$$\text{df} = \frac{(V_1 + V_2)^2}{\dfrac{V_1^2}{n_1 - 1} + \dfrac{V_2^2}{n_2 - 1}} \quad \text{where} \quad V_1 = \frac{s_1^2}{n_1} \quad \text{and} \quad V_2 = \frac{s_2^2}{n_2}$$

df should be truncated (rounded down) to an integer. The t critical values for the usual confidence levels are given in Appendix Table 3.

For a comparison of two treatments, when

1. individuals or objects are randomly assigned to treatments (or vice versa)
2. the sample sizes are large (generally 30 or larger)

   or

   the treatment response distributions are approximately normal, the two-sample t confidence interval formula can be used to estimate $\mu_1 - \mu_2$.

Example 11.4 Freshman Year Weight Gain

Understand the context ➤

● The paper **“Predicting the ‘Freshman 15’: Environmental and Psychological Predictors of Weight Gain in First-Year University Students” (*Health Education Journal* [2010]: 321–332)** described a study conducted by researchers at Carleton University in Canada. The researchers studied a random sample of first-year students who lived on campus and a random sample of first-year students who lived off campus. Data on weight gain (in kg) during the first year, consistent with summary quantities given in the paper, are given on the next page. A negative weight gain represents a weight loss. The researchers believed that the mean weight gain of students living on campus was greater than the mean weight gain for students living off campus and were interested in estimating the difference in means for these two groups using a 95% confidence interval.

● **Data set available online**

On Campus	Off Campus
2.0	1.6
2.3	3.1
1.1	−2.8
−2.0	0.0
−1.9	0.2
5.6	2.9
2.6	−0.9
1.1	3.8
5.6	0.7
8.2	−0.1

For these samples:

On campus: $n_1 = 10 \quad \bar{x}_1 = 2.46 \quad s_1 = 3.26$

Off campus: $n_2 = 10 \quad \bar{x}_2 = 0.85 \quad s_2 = 2.03$

Formulate a plan ➤ You want to estimate

$$\mu_1 - \mu_2 = \text{mean difference in weight gain}$$

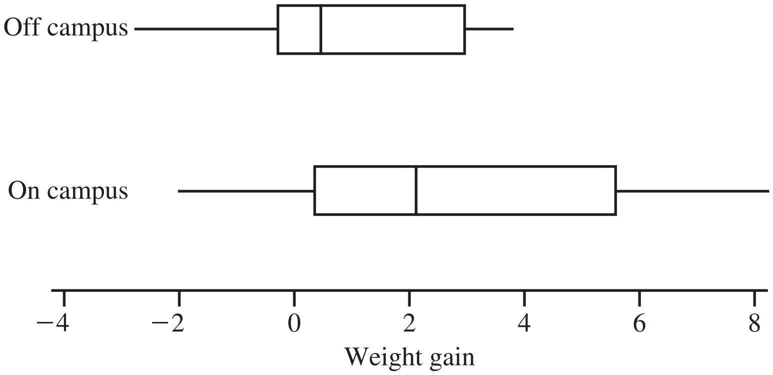
where

$$\mu_1 = \text{mean weight gain for first-year students living on campus}$$

and

$$\mu_2 = \text{mean weight gain for first-year students living off campus}$$

The samples were random samples from the two populations of interest (first-year students who live on campus and first-year students who live off campus), so the samples were independently selected. Because the sample sizes were not large, you need to be willing to assume that the population weight gain distributions are at least approximately normal. Boxplots constructed using the data from the two samples are shown below. There are no outliers in either data set and the boxplots are reasonably symmetric, suggesting that the assumption of approximate normality is appropriate for each of the populations.



Do the work ➤ To estimate $\mu_1 - \mu_2$, the difference in mean weight gain for the two treatments, you can calculate a 95% confidence interval.

$$V_1 = \frac{s_1^2}{n_1} = \frac{(3.26)^2}{10} = 1.06 \qquad V_2 = \frac{s_2^2}{n_2} = \frac{(2.03)^2}{10} = 0.41$$

$$\text{df} = \frac{(V_1 + V_2)^2}{\frac{V_1^2}{n_1 - 1} + \frac{V_2^2}{n_2 - 1}} = \frac{(1.06 + 0.41)^2}{\frac{(1.06)^2}{9} + \frac{(0.41)^2}{9}} = \frac{2.16}{0.14} = 15.43$$

Copyright 2025 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s). Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

Truncating to an integer gives df = 15. In the 15-df row of Appendix Table 3, the *t* critical value for a 95% confidence level is 2.13. The interval is then

$$(\bar{x}_1 - \bar{x}_2) \pm (t \text{ critical value})\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$$

$$= (2.46 - 0.85) \pm (2.13)\sqrt{\frac{(3.26)^2}{10} + \frac{(2.03)^2}{10}}$$

$$= 1.61 \pm (2.13)(1.214)$$

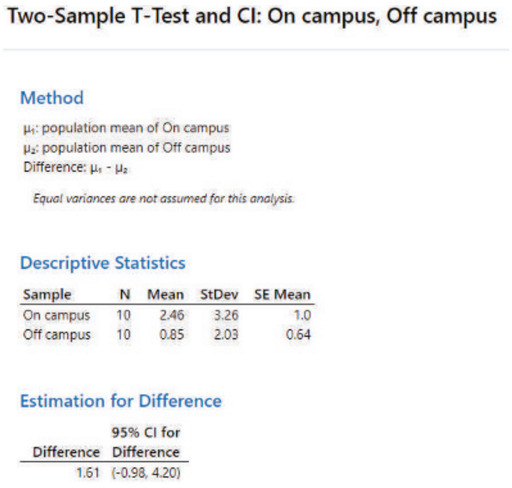
$$= 1.61 \pm 2.586$$

$$= (-0.976, 4.196)$$

Interpret the results ➤ This interval is rather wide, because the two sample sizes are small. Notice that the interval includes 0, so 0 is one of the plausible values for $\mu_1 - \mu_2$. Because 0 is included in this interval, it is possible that there is no real difference in the mean weight gain for students who live on campus and those who live off campus. The 95% confidence level means that you used a method to produce this estimate that correctly captures the actual value of $\mu_1 - \mu_2$ about 95% of the time in repeated sampling.

Suppose that the confidence interval had been (0.976, 4.196), so that the interval did not include 0. This interval would have been interpreted by saying that you are 95% confident that the mean weight gain for those who live on campus is greater than the mean weight gain for those who live off campus by somewhere between 0.976 and 4.196 kg.

Most statistical computer packages can compute the two-sample *t* confidence interval. Minitab was used to construct a 95% confidence interval using the data of this example; the resulting output is shown here:



```
Two-Sample T-Test and CI: On campus, Off campus

Method

μ₁: population mean of On campus
μ₂: population mean of Off campus
Difference: μ₁ - μ₂

  Equal variances are not assumed for this analysis.

Descriptive Statistics

Sample        N   Mean  StDev  SE Mean
On campus    10   2.46   3.26      1.0
Off campus   10   0.85   2.03     0.64

Estimation for Difference

              95% CI for
 Difference   Difference
       1.61  (-0.98, 4.20)
```

Exercises 11.1–11.21

● Data set available online

11.1 Consider two populations for which $\mu_1 = 30$, $\sigma_1 = 2$, $\mu_2 = 25$, and $\sigma_2 = 3$. Suppose that two independent random samples of sizes $n_1 = 40$ and $n_2 = 50$ are selected. Describe the approximate sampling distribution of $\bar{x}_1 - \bar{x}_2$ (center, variability, and shape).

11.2 Descriptions of four studies are given. In each of the studies, the two populations of interest are first-year students at a particular university and second-year students at the university. Which of these studies have samples that are independently selected?

Study 1: To determine if there is evidence that the mean amount of time spent playing online video games differs for the two populations, a random sample of 20 first-year students and a random sample of 50 second-year students are selected.

Study 2: To determine if the mean number of hours worked per week differs for the two populations, a random sample of first-year students is selected. For each of these first-year students, a second-year student who is enrolled in the same number of courses is identified and included in the sample of second-year students.

Study 3: To determine if the mean amount of time spent using campus recreational facilities differs for the two populations, a random sample of first-year students is selected. A separate random sample of the same size is selected from the population of second-year students.

Study 4: To determine if the mean amount of money spent on housing differs for the two populations, a random sample of first-year students is selected. Each student in this sample is asked how far away from campus they live. For each of these first-year students, a second-year student who lives about the same distance from campus is identified and included in the sample of second-year students.

11.3 Many people now turn to the Internet to get information on health-related topics. The paper **"An Examination of Health, Medical and Nutritional Information on the Internet: A Comparative Study of Wikipedia, WebMD and the Mayo Clinic Websites" (*The International Journal of Communication and Health* [2015]: 30–38)** used Flesch reading ease scores (a measure of reading difficulty based on factors such as sentence length and number of syllables in the words used) to score pages on Wikipedia and on WebMD. Higher Flesch scores correspond to more difficult reading levels. The paper reported that for a representative sample of health-related pages on Wikipedia, the mean Flesch score was 26.7 and the standard deviation of the Flesch scores was 14.1. For a representative sample of pages from WebMD, the mean score was 43.9 and the standard deviation was 19.4. Suppose that these means and standard deviations were based on samples of 40 pages from each site. Is there convincing evidence that the mean reading level for health-related pages differs for Wikipedia and WebMD? Test the relevant hypotheses using a significance level of 0.05.

11.4 Use the information in the previous exercise to answer the following questions.

a. Construct a 90% confidence interval estimate of the difference in mean Flesch reading ease score for health-related pages on Wikipedia and health-related pages on WebMD.

b. What does this confidence interval imply about the readability of health-related information from these two sources? Is this consistent with the conclusion in the hypothesis test of the previous exercise?

11.5 A study to investigate differences in risk-taking between people who have dogs and people who have cats is described in the paper **"The Pet Exposure Effect: Exploring the Differential Impact of Dogs Versus Cats on Consumer Mindsets" (*Journal of Marketing* [2022]: 42–57)**. Participants included a sample of 77 dog owners and a sample of 68 cat owners. People who owned both dogs and cats were excluded from the study. Participants were provided with information on mutual funds and stock investments and were told that stock investments had a higher level of risk while mutual funds had a lower level of risk and were a more conservative investment. They were then asked to imagine that they had \$2000 to invest and asked how they would allocate that investment between stocks and mutual funds. The sample means and standard deviations for the amount of money (in dollars) allocated to stock investments are given in the accompanying table.

Sample	n	$\bar{x}$	s
Dog Owners	77	\$796.10	524.45
Cat Owners	68	\$603.69	474.90

a. Both sample standard deviations are quite large. What does this tell you about the distribution of amount of money allocated to stock investments for the two samples?

b. Do the data provide convincing evidence that the mean amount of money allocated to stock investments is greater for dog owners than for cat owners? For purposes of this exercise, you may assume that the samples are representative of the population of dog owners and the population of cat owners. Use a significance level of $\alpha = 0.05$ for your test.

c. The authors of the paper hypothesized that dog owners would exhibit a greater tolerance for risk. Is this supported by the result of the hypothesis test in Part (b)? Explain.

11.6 The authors of the paper **"Facebook and Older Adults: Fulfilling Psychological Needs?" (*Journal of Aging Studies* [2020]: 100897)** explored the ways in which older adults use social media. One question they were interested in was whether Facebook use differed for those who were less mobile than for those who were more mobile. Mobility was defined as "the extent of movement within a person's environment." Study participants were classified into three mobility samples (low, medium, and high) based on frequency of movement and whether they needed help, such as a cane, walker, or personal assistance. Participants in each of the three samples were asked how much time they spent using Facebook in a typical day. Sample means and standard deviations are given in the accompanying table.

Copyright 2025 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s). Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

Sample	n	$\bar{x}$	s
Low Mobility	42	95.55	84.48
Medium Mobility	35	58.31	47.53
High Mobility	32	49.03	48.52

- **a.** You would like to use the data from this study to test hypotheses about the difference in mean time spent on Facebook for the people with low mobility and people with high mobility. What must you be willing to assume about the samples in this study?
- **b.** Carry out a hypothesis test to determine if there is convincing evidence that the mean amount of time spent on Facebook is greater for older adults with low mobility than the mean for older adults with high mobility.

11.7 Use the information given in the previous exercise to determine if there is convincing evidence that the mean amount of time spent on Facebook is different for older adults with medium mobility and older adults with high mobility.

11.8 The paper **"If It's Hard to Read, It's Hard to Do" (*Psychological Science* [2008]: 986–988)** described an interesting study of how people perceive the effort required to do certain tasks. Each of 20 students was randomly assigned to one of two groups. One group was given instructions for an exercise routine that were printed in an easy-to-read font (Arial). The other group received the same set of instructions, but printed in a font that is considered difficult to read (Brush).

After reading the instructions, subjects estimated the time (in minutes) they thought it would take to complete the exercise routine. Summary statistics are given below.

	Easy Font	**Difficult Font**
n	10	10
$\bar{x}$	8.23	15.10
s	5.61	9.28

The authors of the paper used these data to carry out a two-sample t test, and concluded that at the 0.10 significance level, there was convincing evidence that the mean estimated time to complete the exercise routine was less when the instructions were printed in an easy-to-read font than when printed in a difficult-to-read font. Discuss the appropriateness of using a two-sample t test in this situation.

11.9 The authors of the paper introduced in Example 11.1 **("Demographic Implications of Lead Poisoning for Eagles Across North America," *Science* [2022]: 779–782)** collected data on blood lead concentration (in mg/dL) for a representative sample of 57 golden eagles from the Atlantic coast and a representative sample of 214 golden eagles from the Pacific coast. For the sample from the Atlantic coast, $\bar{x}= 26.63$ and $s_1 = 45.08$. For the sample from the Pacific coast, $\bar{x} = 14.35$ and $s_2 = 26.96$. Carry out a test to determine if there is convincing evidence that the mean blood lead concentration is different for Atlantic coast golden eagles and Pacific coast golden eagles. Use a significance level of 0.01 for your test.

11.10 Use the information in the previous exercise to answer the following questions.

- **a.** Construct and interpret a 99% confidence interval for the mean difference in blood lead concentration for Atlantic coast golden eagles and Pacific coast golden eagles.
- **b.** Explain how the confidence interval in Part (a) is consistent with the conclusion reached in the hypothesis test from the previous exercise.

11.11 Just how dirty are the money and credit cards you carry in your wallet? The authors of the article **"Dirty Money: New Research Reveals the Filthiness of Our Cash, Cards, & Coins" (lendedu.com, 2020)** measured the bacteria on the surface of credit cards and money (bills and coins). Germ scores (with higher scores corresponding to dirtier surfaces) for 41 credit cards and 39 bills and coins were determined. For the sample of credit cards, the mean germ score was 285 and the standard deviation was 290. For the sample of bills and coins, the mean germ score was 152 and the standard deviation was 144. For purposes of this exercise, you can assume that it is reasonable to regard these two samples as representative of credit cards and money.

- **a.** The standard deviation is greater than the mean for the sample of credit card germ scores and is nearly equal to the mean for the sample of money germ scores. Explain why it is not reasonable to think that the distribution of germ scores would be approximately normal for either credit cards or money.
- **b.** Given your response to Part (a), would it be appropriate to use the two-sample t test to test the null hypothesis that there is no difference in the mean germ score for credit cards and money? Explain why or why not.
- **c.** If appropriate, carry out a test to determine if there is convincing evidence that the mean germ score is greater for credit cards than for money. Use $\alpha = 0.05$.

11.12 The paper **"Mood Food: Chocolate and Depressive Symptoms in a Cross-Sectional Analysis" (*Archives of Internal Medicine* [2010]: 699–703)** describes a study that investigated the relationship between depression and chocolate consumption. Participants in the study were 931 adults who were not currently taking medication for depression. These participants were screened for depression using a widely used screening test. The participants were then divided into two samples based on the score on the screening test. One sample consisted of people who screened positive for depression, and the other sample consisted of people who did not screen positive for depression. Each of the study participants also completed a food frequency survey.

The researchers believed that the two samples were representative of the two populations of interest—adults who would screen positive for depression and adults who would not screen positive. The paper reported that the mean number of servings of chocolate for the sample of people that screened positive for depression was 8.39 servings per month and the sample standard deviation was 14.83. For the sample of people who did not screen positive for depression, the mean number of servings per month was 5.39 and the standard deviation was 8.76. The paper did not say how many individuals were in each sample, but for the purposes of this exercise, you can assume that the 931 study participants were divided into 311 who screened positive for depression and 620 who did not screen positive.

Estimate the difference in the mean number of servings of chocolate per month in the population of people who would screen positive for depression and the mean number of chocolate servings per month in the population of people who would not screen positive for depression. Use a confidence level of 90% and be sure to interpret the interval in context.

11.13 ● The authors of the paper **"Influence of Biofeedback Weight Bearing Training in Sit to Stand to Sit and the Limits of Stability on Stroke Patients" (*The Journal of Physical Therapy Science* [2016]: 3011–3014)** randomly selected two samples of patients admitted to the hospital after suffering a stroke. One sample was selected from patients who received biofeedback weight training for 8 weeks, and the other sample was selected from patients who did not receive this training. At the end of 8 weeks, the time it took (in seconds) to stand from a sitting position and then to sit down again (called sit-stand-sit time) was measured for the people in each sample. The accompanying data are consistent with summary quantities given in the paper. For purposes of this exercise, you can assume that the samples are representative of the population of stroke patients who receive the biofeedback training and the population of stroke patients who do not receive this training. Use the given data to construct and interpret a 95% confidence interval for the difference in mean sit-stand-sit time for these two populations.

Biofeedback Group

1.9 2.6 4.3 2.1 2.7 4.1 3.2 4.0 3.2 3.5 2.8 3.5 3.5 2.3 3.1

No Biofeedback Group

5.1 4.7 3.9 4.2 4.7 4.3 4.2 5.1 3.4 4.2 5.1 4.4 4.0 3.4 3.9

11.14 ● Do long periods of social isolation impact the brain? The authors of the paper **"Brain Changes in Response to Long Antarctic Expeditions" (*New England Journal of Medicine* [2019]: 381)** investigated by measuring changes in regions in the brain for a sample of eight adults who were socially isolated during a 12-month expedition to Antarctica and a sample of nine adults who served as controls and who did not spend time in Antarctica. Changes were calculated as the difference in volume at the end and at the beginning of the study, so a negative value represents a decrease in volume. For a region of the brain called the Dentate Gyrus (DG), the changes in volume (mm^3) were:

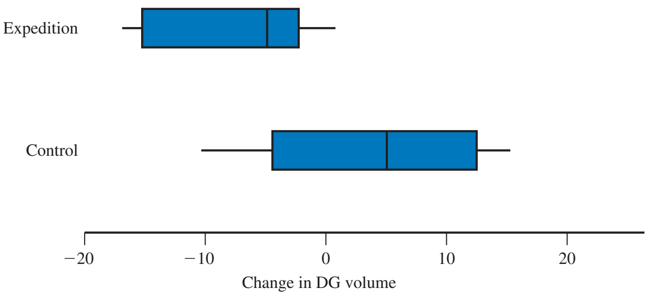
Expedition: −16.29, −15.90, −13.34, −6.16, −3.52, −2.23, −2.22, 0.91

Control: −10.31, −5.41, −3.40, −1.56, 5.11, 6.96, 12.37, 12.84, 15.31

a. Why do you think it was important to include the control sample in this study?

b. If you wanted to use these data to determine if the mean change in DG volume was different for adults who spend a year isolated in Antarctica and adults who do not, what would you need to assume about the samples used in this study?

c. One of the conditions needed to use the two-sample *t* test with small samples is that it is reasonable to assume that the population distributions are approximately normal. Below are boxplots of the sample data. Based on these boxplots, is reasonable to make this assumption? Provide a justification for your answer.



d. Do these data provide convincing evidence that the mean change in DG volume is different for adults who spend a year isolated in Antarctica and adults who do not? Test the relevant hypotheses using a significance level of $\alpha = 0.05$.

11.15 Would high school students get more sleep if school had a later start time? This question is investigated

Copyright 2025 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s). Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

in the paper **"The Association Between School Start Time and Sleep Duration, Sustained Attention, and Academic Performance" (*Nature and Science of Sleep* [2020]: 1161–1172).** Each student in a sample 21 students with a late start (9:00) and a sample of 30 students with an early start (8:00) were asked to keep a sleep diary to record their length of sleep. The authors reported that the average amount of sleep on school days for those in the late start sample was 475 minutes and it was 450 minutes for those in the early start sample. Sample standard deviations (estimated from graphs that appeared in the paper) were 11 minutes for the late start sample and 14 minutes for the early start sample.

a. The authors of the paper used a two-sample *t* test to conclude that the mean length of sleep on school nights was greater for students with the late school start time. What assumptions about the samples would they have had to make in order for the two-sample *t* test to be an appropriate hypothesis testing method?

b. Assume that the necessary conditions are met. Carry out a test to determine if the mean length of sleep on a school night is greater for students with a late school start time than for students with an early school start time.

c. In this study, participants were asked to self-select whether they wanted to be in the late start time sample or the early start time sample. Discuss the implications of this aspect of the study design.

11.16 The paper **"Does the Color of the Mug Influence the Taste of the Coffee?" (*Flavour* [2014]: 1–7)** describes an experiment in which subjects were assigned at random to one of two treatment groups. The 12 people in one group were served coffee in a white mug and were asked to rate the quality of the coffee on a scale from 0 to 100. The 12 people in the second group were served the same coffee in a clear glass mug, and they also rated the coffee. The mean quality rating for the 12 people in the white mug group was 50.35 and the standard deviation was 20.17. The mean quality rating for the 12 people in the clear glass mug group was 61.48 and the standard deviation was 16.69. For this exercise, you may assume that the distribution of quality ratings for each of the two treatments is approximately normal.

a. Use the given information to construct and interpret a 95% confidence interval for the difference in mean quality rating for this coffee when served in a white mug and when served in a glass mug.

b. Based on the interval from Part (a), are you convinced that the color of the mug makes a difference in terms of mean quality rating? Explain.

11.17 The paper **"Percolating Ideas: The Effects of Caffeine on Creative Thinking and Problem-Solving" (*Consciousness and Cognition* [2020]: 102899)** describes an experiment in which 88 adults were assigned at random to one of two groups—a caffeine group and a control group. People in the caffeine group consumed a capsule containing 200 mg of caffeine. People in the control group consumed a capsule that looked the same as the caffeine capsule but did not contain any caffeine. After 20 minutes, a measure of working memory was recorded. The researchers reported that there was no significant difference in mean score on the measure of working memory for the two groups. In the context of this experiment, explain what it means to say that there is *no significant difference* in the group means.

11.18 The paper **"Television Advertisements for High-Sugar Foods and Beverages: Effect on Children's Snack Food Intake" (*British Journal of Nutrition* [2021]: 591–597)** describes an experiment in which 101 children were assigned at random to one of two groups. Children in the "sugar" group were shown cartoons that embedded advertisements for high sugar foods. Children in the "toy" group were shown the same cartoons but the embedded advertisements were for toys rather than for high sugar foods. The children were then offered a variety of snacks and beverages and were given 15 minutes to eat or drink as much or as little as they wanted. The researchers reported that children in the sugar group consumed a significantly greater mean number of calories than children in the toy group. In the context of this experiment, explain what it means to say that the mean number of calories consumed was *significantly greater* for the sugar group.

11.19 Fumonisins are environmental toxins produced by a type of mold and have been found in corn and in products made from raw corn. The **Center for Food Safety and Applied Nutrition** provided recommendations on allowable fumonisin levels in human food and in animal feed based on a study of corn meal. The study compared corn meal made from partially degermed corn (corn that has had the germ, the part of the kernel located at the bottom center of the kernel that is used to produce corn oil, partially removed) and corn meal made from corn that has not been degermed. Specimens of corn meal were analyzed and the total fumonisin level (ppm) was determined for each specimen. Summary statistics for total fumonisin level from the U.S. Food and Drug Administration's website are given here.

	$\bar{x}$	s
Partially Degermed	0.59	1.01
Not Degermed	1.21	1.71

Copyright 2025 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s). Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

a. If the given means and standard deviations had been based on a random sample of 10 partially degermed specimens and a random sample of 10 specimens made from corn that was not degermed, explain why it would not be appropriate to carry out a two-sample *t* test to determine if there is a significant difference in the mean fumonisin level for the two types of corn meal.

b. Suppose instead that each of the random samples had included 50 corn meal specimens. Explain why it would now be reasonable to carry out a two-sample *t* test.

c. Assuming that each random sample size was 50, carry out a test to determine if there is a significant difference in mean fumonisin level for the two types of corn meal. Use a significance level of 0.01.

11.20 A researcher at the Medical College of Virginia conducted a study of 60 randomly selected soccer players and concluded that frequently "heading" the ball in soccer lowers players' IQs (***USA TODAY,* August 14, 1995**). The soccer players were divided into two groups, based on whether they averaged 10 or more headers per game.

Mean IQs were reported in the article, but the sample sizes and standard deviations were not given. Suppose that these values were as given in the accompanying table.

	n	Sample Mean	Sample SD
Fewer Than 10Headers	35	112	10
10or More Headers	25	103	8

a. Do these data support the researcher's conclusion? Test the relevant hypotheses using $\alpha = 0.05$.

b. Can you conclude that heading the ball *causes* lower IQ?

11.21 Do certain behaviors result in a severe drain on energy resources because a great deal of energy is expended in comparison to energy intake? The article **"The Energetic Cost of Courtship and Aggression in a Plethodontid Salamander" (*Ecology* [1983]: 979–983)** reported on one of the few studies concerned with behavior and energy expenditure. The accompanying table gives oxygen consumption (mL/g/hr) for male-female salamander pairs.

Behavior	Sample Size	Sample Mean	Sample SD
Noncourting	11	0.072	0.0066
Courting	15	0.099	0.0071

a. The pooled *t* test is a test procedure for testing $H_0\colon \mu_1 - \mu_2 = \textit{hypothesized value}$ when it is reasonable to assume that the two population distributions are normal with equal standard deviations ($\sigma_1 = \sigma_2$). The test statistic for the pooled *t* test is obtained by replacing both s_1 and s_2 in the two-sample *t* test statistic with s_p where

$$s_p = \sqrt{\frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}}$$

When the population distributions are normal with equal standard deviations and H_0 is true, the resulting pooled *t* statistic has a *t* distribution with $\text{df} = n_1 + n_2 - 2$. For the reported data, the two sample standard deviations are similar. Use the pooled *t* test with $\alpha = 0.05$ to determine whether the mean oxygen consumption for courting pairs is higher than the mean oxygen consumption for noncourting pairs.

b. Would the conclusion in Part (a) have been different if the two-sample *t* test had been used rather than the pooled *t* test?

Section 11.2 Inferences Concerning the Difference Between Two Population or Treatment Means Using Paired Samples

Two samples are said to be **independent** if the selection of the individuals or objects that make up one of the samples has no bearing on the selection of individuals or objects in the other sample. In some situations, a study with independent samples is not the best way to obtain information about a possible difference between two populations. For example, suppose that an investigator wants to determine whether regular aerobic exercise affects blood pressure. A random sample of people who jog regularly and a second random sample of people who do not exercise regularly are selected independently of one another.

The researcher might use the two-sample *t* test to conclude that a significant difference exists between the mean blood pressures for joggers and nonjoggers. But is it reasonable

Copyright 2025 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s). Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

to think that the difference in mean blood pressure is the result of jogging? It is known that blood pressure is related to both diet and body weight. Maybe the joggers in the sample tend to be leaner and adhere to a healthier diet than the nonjoggers. This might account for the observed difference. On the basis of this study, the researcher would not be able to rule out the possibility that the observed difference in blood pressure is explained by weight differences between the people in the two samples and that aerobic exercise itself has no effect.

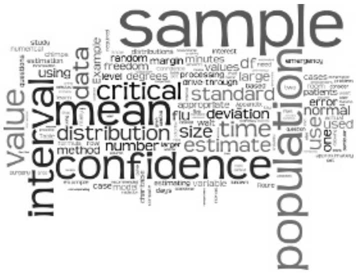
One way to avoid this difficulty is to match subjects by weight. The researcher could find pairs of subjects where the jogger and nonjogger in each pair were similar in weight (although weights for different pairs might vary widely). The factor *weight* could then be ruled out as a possible explanation for an observed difference in mean blood pressure between the two groups. Matching the subjects by weight results in two samples for which each observation in the first sample is paired in a meaningful way with a particular observation in the second sample. Such samples are said to be **paired**.

Paired samples occur in a number of different ways. Some studies involve using the same group of individuals with measurements recorded both before and after some intervening treatment. Other studies might use naturally occurring pairs, such as twins or married couples. Others construct pairs by matching on factors with effects that might otherwise make differences (or the lack of them) between the two populations difficult to detect (such as weight in the jogging example).

Example 11.5 illustrates why it is important to consider whether two samples are independent or paired.

Example 11.5 Document Word Clouds

Understand the context ➤



Source: wordle.net

● When people are searching for information online, most people quickly scan a document and attempt to determine if it is relevant before taking the time to read the whole document. The authors of the paper **"Document Word Clouds: Visualising Web Documents as Tag Clouds to Aid Users in Relevance Decisions" (*Research and Advanced Technology for Digital Libraries* [2009]: 94–105)** wondered if people would have an easier time determining if a document was relevant if they first saw a word cloud representation of the document. (A word cloud is a visual representation of a text passage that uses characteristics such as font size, color, and color density to indicate the importance of a particular word.)

The researchers chose 10 documents and created a word cloud representation for each of them. The text document version was shown to one group of people and the word cloud representation was shown to another group of people. The average time (in seconds) it took the people in the group to make a relevance decision was recorded for each version of the 10 documents is shown in the accompanying table.

Document	Time to Relevance Decision Text Version	Time to Relevance Decision Word Cloud Version
1	3.55	2.95
2	2.94	3.04
3	3.72	2.72
4	2.63	1.97
5	2.39	2.17
6	5.36	3.64
7	1.86	1.96
8	2.49	2.04
9	2.76	2.46
10	2.77	2.63

You can view these data as consisting of two samples—a sample of times for the relevance decision for text documents and a sample of times for the relevance decision for word cloud representations.

In this case, the samples are paired rather than independent because each time in the text document sample is paired with a particular time in the word cloud sample for the same document. Notice that in 8 of the 10 data pairs, the time to relevance decision is greater for

● **Data set available online**

Copyright 2025 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s). Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

the text version of the document than for the word cloud representation of the same document. This suggests that the two population means may be different.

Both the text document times and the word cloud version times vary from document to document. It is this variability that may make any difference difficult to see if the paired nature of the samples is ignored. To see how this might be the case, consider the two plots in Figure 11.1. The first plot (Figure 11.1 (a)) ignores the pairing, and the two samples look quite similar. However, the plot in which pairs are identified (Figure 11.1 (b)) does suggest that there is a difference because for eight of the ten pairs the text document time to relevance decision is greater than the time for the word cloud representation (these are the pairs linked by line segments that slant to the right). And in the two cases where the time for the text document is less than the time for the word cloud, the line segments are close to vertical, indicating that the two times for those documents were not very different.

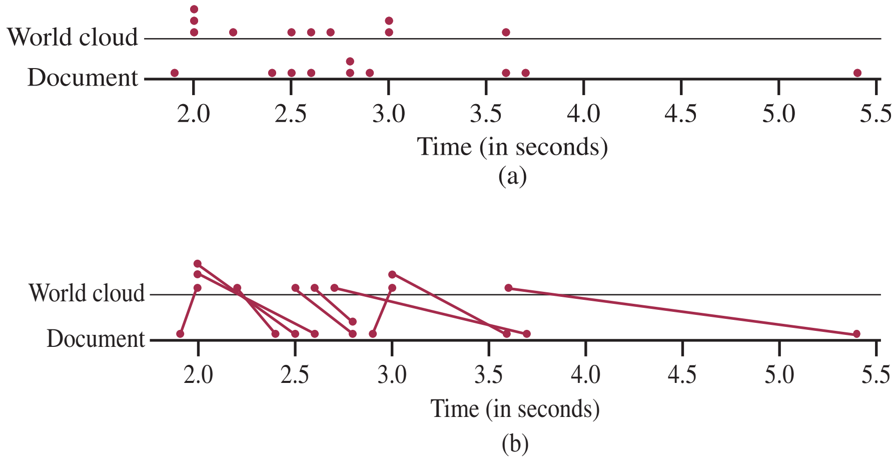


Figure 11.1 Two plots of the paired data from Example 11.5: (a) pairing ignored; (b) pairs identified.

Example 11.5 suggests that the methods of inference developed for independent samples are not adequate for dealing with paired samples. When sample observations from the first population are paired in some meaningful way with sample observations from the second population, inferences should be based on the differences between the two observations within each sample pair. The n sample differences can then be regarded as having been selected from a large population of differences. For example, in Example 11.5, you can think of the 10 (text − word cloud) differences as having been selected from an entire population of differences (corresponding to a population of documents).

Before considering how the sample of differences is used to test hypotheses about the difference in population means, some new notation is needed:

$$\mu_d = \text{mean value of the difference population}$$

and

$$\sigma_d = \text{standard deviation of the difference population}$$

The relationship between the two individual population means and the mean difference is

$$\mu_d = \mu_1 - \mu_2$$

This means that when the samples are paired, inferences about $\mu_1 - \mu_2$ are equivalent to inferences about μ_d. Since inferences about μ_d can be based on the n observed sample differences, the original two-sample problem becomes a familiar one-sample problem.

Paired *t* Test

To compare two population or treatment means when the samples are paired, first translate the hypotheses of interest from ones about the value of $\mu_1 - \mu_2$ to equivalent hypotheses involving μ_d:

Copyright 2025 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s). Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

Hypothesis	**Equivalent Hypothesis When Samples Are Paired**
H_0: $\mu_1 - \mu_2 =$ hypothesized value	H_0: $\mu_d =$ hypothesized value
H_a: $\mu_1 - \mu_2 >$ hypothesized value	H_a: $\mu_d >$ hypothesized value
H_a: $\mu_1 - \mu_2 <$ hypothesized value	H_a: $\mu_d <$ hypothesized value
H_a: $\mu_1 - \mu_2 \neq$ hypothesized value	H_a: $\mu_d \neq$ hypothesized value

The general form of the null hypothesis is

$$H_0\text{: } \mu_1 - \mu_2 = \text{hypothesized value}$$

In most cases, the hypothesized value is 0, indicating that there is no difference in population means ($\mu_1 = \mu_2$). But sometimes the hypothesized value is something other than 0. For example, there are two different routes between San Luis Obispo, California, and Monterey, California. One is an inland route and the other is a more scenic coastal route. The scenic route is known to usually take more time, but you might be interested in deciding if the mean travel time for the scenic route is more than 1 hour longer. Because travel time can depend on day of the week and time of day, suppose that you collect data by having two drivers leave at the same time, with one taking the inland route and the other taking the coastal route. You repeat this process on 11 more days to get a sample of 12 travel times for each of the two routes.

With μ_1 representing the mean travel time for the scenic route and μ_2 representing the mean travel time for the inland route, the data from these two samples, which are paired by the day and time of the trip, can be used to test

$$H_0\text{: } \mu_1 - \mu_2 = 1 \text{ versus } H_0\text{: } \mu_1 - \mu_2 > 1$$

which is equivalent to

$$H_0\text{: } \mu_d = 1 \text{ versus } H_0\text{: } \mu_d > 1$$

Sample differences (Sample 1 value − Sample 2 value) are calculated and used as the basis for testing hypotheses about μ_d. When the number of differences is large or when it is reasonable to assume that the population of differences is approximately normal, the one-sample *t* test based on the differences is the recommended test procedure. In general, if each of the two individual populations is normal, the population of differences is also normal. A normal probability plot or boxplot of the differences can be used to decide if the assumption of normality is reasonable.

Paired *t* Test for Comparing Two Population or Treatment Means

Null Hypothesis: H_0: $\mu_d =$ hypothesized value

Test Statistic: $$t = \frac{\bar{x}_d - \text{hypothesized value}}{\frac{s_d}{\sqrt{n}}}$$

where n is the number of sample differences and $\bar{x}_d$ and s_d are the mean and standard deviation of the sample differences. This test is based on df $= n - 1$.

Alternative Hypothesis:	***P*-Value:**
H_a: $\mu_d >$ hypothesized value	Area under the appropriate *t* curve to the right of *t*
H_a: $\mu_d <$ hypothesized value	Area under the appropriate *t* curve to the left of *t*
H_a: $\mu_d \neq$ hypothesized value	2(area to the right of *t*) if *t* is positive or 2(area to the left of *t*) if *t* is negative

Assumptions:

1. The samples are paired.
2. The *n* sample differences can be viewed as a random sample from a population of differences.
3. The number of sample differences is large (generally at least 30) or the population distribution of differences is approximately normal.

Example 11.6 Word Clouds Revisited

● The time to relevance decision data of Example 11.5 can be used to test the claim that the mean time is greater for text documents than for word cloud representations. Because the samples are paired, the first thing to do is calculate the sample differences. These are the text – word cloud time differences for the 10 documents in the sample.

The sample data and the calculated differences are shown in the accompanying table. A positive difference indicates that the time to make a relevance decision was greater for the text document than for the word cloud representation.

Document	Time (in seconds) to Relevance Decision Text Version	Time (in seconds) to Relevance Decision Word Cloud Version	Difference (Text – Word Cloud)
1	3.55	2.95	0.60
2	2.94	3.04	−0.10
3	3.72	2.72	1.00
4	2.63	1.97	0.66
5	2.39	2.17	0.22
6	5.36	3.64	1.72
7	1.86	1.96	−0.10
8	2.49	2.04	0.45
9	2.76	2.46	0.30
10	2.77	2.63	0.14

For this example,

μ_1 = mean time to make a relevance decision for text documents
μ_2 = mean time to make a relevance decision for word cloud representations

and

$\mu_d = \mu_1 - \mu_2$ = mean difference in time (text − word cloud)

The question of interest can be answered by testing the hypothesis

$$H_0\colon\ \mu_d = 0 \qquad \text{versus} \qquad H_a\colon\ \mu_d > 0$$

The mean and standard deviation of the sample differences are $\bar{x}_d = 0.489$ and $s_d = 0.552$. Do these data provide evidence that the mean time to make a relevance decision is greater for text documents than for word cloud representations of documents?

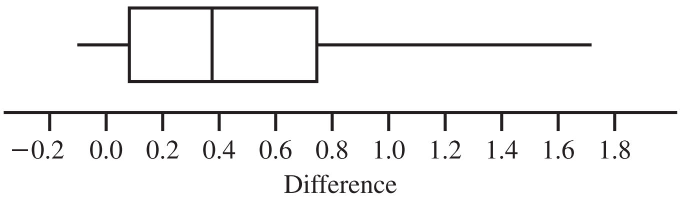
The paired *t* test with a significance level of 0.05 will be used to carry out the hypothesis test.

Understand the context ➤

1. μ_d = mean difference in time (text − word cloud)
2. $H_0\colon\ \mu_d = 0$
3. $H_a\colon\ \mu_d > 0$
4. Significance level: $\alpha = 0.05$

Formulate a plan ➤

5. Test statistic: $t = \dfrac{\bar{x}_d - \text{hypothesized value}}{\dfrac{s_d}{\sqrt{n}}}$
6. Assumptions: Although the sample of 10 documents was not a random sample, the researchers selected the documents to be representative of the population of documents. For this reason, it is reasonable to view the 10 sample differences as a random sample of all such differences. A boxplot of the differences is not too asymmetric and does not show any outliers, so the assumption of normality is not unreasonable.



● **Data set available online**

Copyright 2025 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s). Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

Do the work ➤

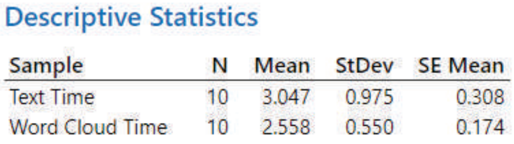
7. Calculation: $t = \dfrac{0.489 - 0}{\dfrac{0.552}{\sqrt{10}}} = 2.79$

8. P-value: This is an upper-tailed test, so the P-value is the area to the right of the calculated t value. The appropriate df for this test is df $= 10 - 1 = 9$. From the 9-df column of Appendix Table 4, P-value $=$ area to the right of $2.79 \approx 0.010$.

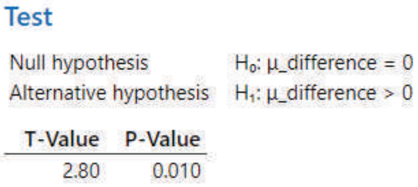
Interpret the results ➤

9. Conclusion: Because the P-value (0.010) is less than α (0.05), you reject H_0. There is convincing evidence that the mean time to make a relevance decision is greater for text documents than the mean time for word cloud representations.

Statistical software could also be used to calculate the value of the test statistic, degrees of freedom, and the P-value. Minitab output is shown here:



```
Descriptive Statistics

Sample            N   Mean   StDev   SE Mean
Text Time        10   3.047  0.975     0.308
Word Cloud Time  10   2.558  0.550     0.174
```



```
Test

Null hypothesis          H₀: μ_difference = 0
Alternative hypothesis   H₁: μ_difference > 0

  T-Value   P-Value
     2.80     0.010
```

Using the two-sample t test (for independent samples) for the data in Example 11.6 would have been incorrect, because the samples are not independent. Inappropriate use of the two-sample t test would have resulted in a calculated test statistic value of 1.38. The conclusion would have been to not reject the hypothesis of equal mean times. This illustrates why it is important to recognize when samples are paired.

Example 11.7 Blue Light from Electronic Devices

Smartphones, flat-screen TVs, and other electronic devices emit blue light that might be affecting your sleep. To investigate, the authors of the paper **"Attenuation of Short Wavelengths Alters Sleep and the ipRGC Pupil Response" (*Ophthalmic & Physiological Optics* [2017]: 440–450)** carried out a study in which 21 participants wore glasses that blocked blue light for about 4 hours prior to bedtime for 2 weeks. Nighttime melatonin (a hormone that plays a role in sleep) was measured at the beginning of the study and again at the end of the 2-week period when the glasses were worn.

What Do You Think?

The study described in this example resulted in two samples of melatonin levels—one consisting of measurements before wearing blue light blocking glasses and one consisting of measurements after wearing the glasses before bedtime for 2 weeks. Are these samples paired or independent?

The authors believed that if melatonin levels tended to increase when the exposure to blue light prior to bedtime was limited, it would be evidence that this type of light exposure could reduce the quality of sleep. To test this theory, they decided to carry out a test to determine if there was evidence that the mean melatonin level was greater after the use of the glasses.

The following sample statistics were reported for the differences in melatonin levels (before minus after):

$$n = 21 \qquad \bar{x}_d = -9.4 \qquad s_d = 10.9$$

Copyright 2025 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s). Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

Most statistical software packages will perform a paired *t* test, and Minitab will be used to carry out a test to determine if there is convincing evidence that the mean melatonin level is greater after the use of blue light blocking glasses.

Understand the context ➤

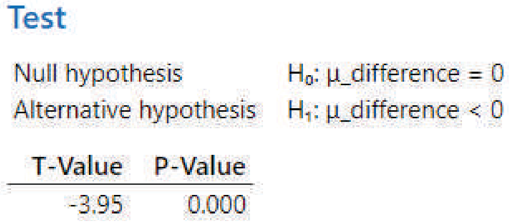
1. μ_d = difference mean melatonin level before wearing blue light blocking glasses and after wearing blue light blocking glasses for 2 weeks
2. H_0: $\mu_d = 0$
3. H_a: $\mu_d < 0$
4. Significance level: $\alpha = 0.01$

Formulate a plan ➤

5. Test statistic: $t = \dfrac{\bar{x}_d - \text{hypothesized value}}{\dfrac{s_d}{\sqrt{n}}}$
6. Assumptions: The authors of the paper believed that the participants in the study could be regarded as representative of people of age 17 to 42. The actual data was not given in the paper, but the authors believed that the conditions for the paired *t* test were met.

Do the work ➤

7. Calculation: From the given Minitab output, $t = -3.95$.



```
Test

Null hypothesis         H₀: μ_difference = 0
Alternative hypothesis  H₁: μ_difference < 0

T-Value  P-Value
-3.95    0.000
```

8. *P*-value: From the Minitab output, *P*-value = 0.000.

Interpret the results ➤

9. Conclusion: The *P*-value (0.000) is less than α (0.01), so the null hypothesis is rejected. There is convincing evidence that the mean melatonin level is greater after wearing blue light blocking glasses for 2 weeks, confirming the researchers' hypothesis that exposure to blue light affects melatonin level and so may have a negative effect on sleep.

Notice that the numerator of the paired *t* test statistic $\bar{x}_d$, and the numerator of the two-sample *t* test statistic, $\bar{x}_1 - \bar{x}_2$, are always equal. The difference is in the denominator. The variability in differences is usually smaller than the variability in each sample separately (because measurements in a pair tend to be similar). As a result, the value of the paired *t* test statistic is usually greater in magnitude than the value of the two-sample *t* test statistic. Pairing usually reduces variability, making differences easier to detect.

A Confidence Interval

The one-sample *t* confidence interval for μ given in Chapter 9 is easily adapted to obtain an interval estimate for μ_d.

Paired *t* Confidence Interval for μ_d

When

1. the samples are paired.
2. the *n* sample differences can be viewed as a random sample from a population of differences.
3. the number of sample differences is large (generally at least 30) or the population distribution of differences is approximately normal,

the paired *t* confidence interval for μ_d is

$$\bar{x}_d \pm (t \text{ critical value}) \cdot \frac{s_d}{\sqrt{n}}$$

For a specified confidence level, the $(n - 1)$ df row of Appendix Table 3 gives the appropriate *t* critical values.

Copyright 2025 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s). Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

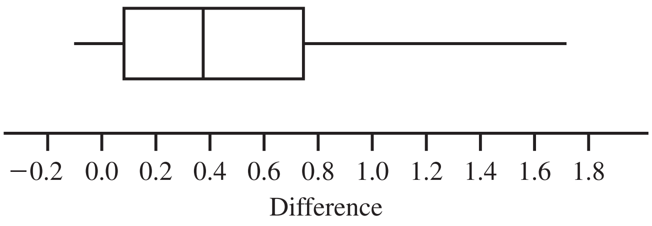
Example 11.8 Word Clouds One Last Time

● Example 11.5 provided data from a study to assess whether the time it takes to make a decision about the relevance of an online document is greater when the document is in text form than when it is represented by a word cloud. The conclusion in the hypothesis test in Example 11.6 was that there is convincing evidence that the mean time to make a relevance decision is greater for text documents than for word cloud representations. Once you have reached this conclusion, it may also be of interest to estimate how much greater the mean time is for text documents than for word clouds.

The sample data and calculated differences from Example 11.6 are shown again in the accompanying table.

Document	Time (in seconds) to Relevance Decision Text Version	Time (in seconds) to Relevance Decision Word Cloud Version	Difference (Text − Word Cloud)
1	3.55	2.95	0.60
2	2.94	3.04	−0.10
3	3.72	2.72	1.00
4	2.63	1.97	0.66
5	2.39	2.17	0.22
6	5.36	3.64	1.72
7	1.86	1.96	−0.10
8	2.49	2.04	0.45
9	2.76	2.46	0.30
10	2.77	2.63	0.14

Formulate a plan ➤ These data will be used to estimate the mean difference in the time it takes to make a relevance decision using a 95% confidence interval. The accompanying boxplot of the 10 sample differences is not too asymmetric and there are no outliers, so it is reasonable to think that the difference population is approximately normal, and the paired *t* confidence interval is appropriate.



Do the work ➤ The mean and standard deviation calculated using the 10 sample differences are $\bar{x}_d = 0.489$ and $s_d = 0.552$. The *t* critical value for df = 9 and a 95% confidence level is 2.26. Substituting these values into the formula for the paired *t* confidence interval gives

$$= 0.489 \pm (2.26)\frac{0.552}{\sqrt{10}}$$

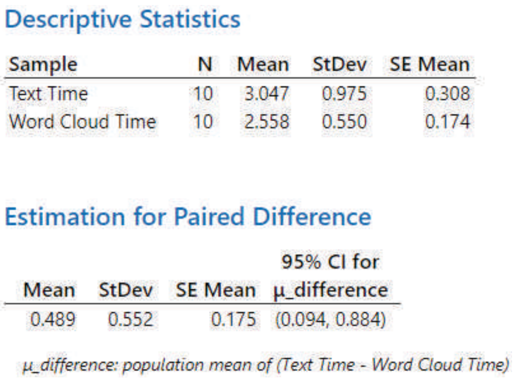
$$= 0.489 \pm 0.395$$

$$= (0.094, 0.884)$$

interpret the results ➤ Based on the sample data, you can be 95% confident that the actual difference in mean time to make a relevance decision is somewhere between 0.094 and 0.884 seconds. Because both endpoints of the interval are positive, this means you would estimate that the mean time to make a relevance decision is greater for text documents than for word cloud representations by somewhere between 0.094 and 0.884 seconds. The method used to construct this interval is successful in capturing the actual difference in population means approximately 95% of the time. Notice that the estimated difference is small, so while the difference is statistically significant, it may not be of any practical significance.

● **Data set available online**

A statistical software package or a graphing calculator could also have been used to determine the endpoints of the confidence interval. Minitab output is shown below.



```
Descriptive Statistics

Sample           N   Mean   StDev  SE Mean
Text Time        10  3.047  0.975    0.308
Word Cloud Time  10  2.558  0.550    0.174


Estimation for Paired Difference

                           95% CI for
 Mean   StDev  SE Mean   μ_difference
 0.489  0.552    0.175   (0.094, 0.884)

 μ_difference: population mean of (Text Time - Word Cloud Time)
```

When two populations must be compared to draw a conclusion on the basis of sample data, a researcher might choose to use independent samples or paired samples. In many situations, paired data allow for a better comparison by screening out the effects of extraneous variables that might make differences between the two populations difficult to detect or that might suggest a difference when none exists.

Exercises 11.22–11.40

● **Data set available online**

11.22 Suppose that you were interested in investigating the effect of a drug that is to be used in the treatment of patients who have glaucoma in both eyes. A comparison between the mean reduction in eye pressure for this drug and for a standard treatment is desired. Both treatments are applied directly to the eye.

a. Describe how you would go about collecting data for your investigation.

b. Does your method result in paired data?

c. Can you think of a reasonable method of collecting data that would not result in paired samples? Would such an experiment be as informative as a paired experiment? Comment.

11.23 ● Many runners believe that listening to music while running enhances their performance. The authors of the paper **"Effects of Synchronous Music on Treadmill Running Among Elite Triathletes" (*Journal of Science and Medicine in Sport* [2012]: 52–57)** wondered if this was true for experienced runners. They recorded time to exhaustion for 11 triathletes while running on a treadmill at a speed determined to be near their peak running velocity. The time to exhaustion was recorded for each participant on two different days. On one day, each participant ran while listening to music that the runner selected as motivational. On a different day, each participant ran with no music playing.

You can assume that it is reasonable to regard these 11 triathletes as representative of the population of experienced triathletes. Only summary quantities were given in the paper, but the data in the table below are consistent with the means and standard deviations given in the paper. Do the data provide convincing evidence that the mean time to exhaustion for experienced triathletes is greater when they run while listening to motivational music? Test the relevant hypotheses using a significance level of 0.05.

Time to Exhaustion (seconds)

Runner	1	2	3	4	5	6	7	8	9	10	11
Motivational Music	535	533	527	524	431	498	555	396	539	542	523
No Music	467	446	482	573	562	592	473	496	552	500	524

11.24 The study described in the previous exercise also measured time to exhaustion for the 11 triathletes on a day when they listened to music that the runners had classified as neutral as compared to motivational. The researchers calculated the difference between the time to exhaustion while running to motivational music and while running to neutral music. The mean difference (motivational − neutral) was −7 seconds (the sample mean time to exhaustion was actually lower when listening to music the runner viewed as motivational than when listening to music the runner viewed as neutral). Suppose that the standard deviation of the differences was $s_d = 80$. You can assume that it is reasonable to regard these 11 triathletes as representative of the population of experienced triathletes and that the population difference distribution is approximately normal. Is there convincing evidence that the

Copyright 2025 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s). Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

mean time to exhaustion for experienced triathletes running to motivational music differs from the mean time to exhaustion when running to neutral music? Carry out a hypothesis test using $\alpha = 0.05$.

11.25 In Exercise 11.23, a hypothesis test leads to the conclusion that there is not convincing evidence that the mean time to exhaustion for experienced triathletes is greater when they run while listening to motivational music. Use the information given in that exercise to construct and interpret a 95% confidence interval for the difference in mean time to exhaustion for experienced triathletes when running to motivational music and the mean time when running with no music.

11.26 The article **"Puppy Love? It's Real, Study Says" (*USA TODAY*, April 17, 2015)** describes a study into how people communicate with their pets. The conclusion expressed in the title of the article was based on research published in ***Science* ("Oxytocin-Gaze Positive Loop and the Coevolution of Human-Dog Bonds," April 17, 2015)**. Researchers measured the oxytocin levels (in picograms per milligram, pg/mg) of 22 dog owners before and again after a 30-minute interaction with their dogs. (Oxytocin is a hormone known to play a role in parent-child bonding.) The difference in oxytocin level (after − before)) was calculated for each of the 22 dog owners. Suppose that the mean and standard deviation of the differences (approximate values based on a graph in the paper) were $\bar{x}_d = 27$ pg/mg and $s_d = 30$ pg/mg.

a. Explain why the two samples (oxytocin levels before interaction and oxytocin levels after interaction) are paired.

b. Assume that it is reasonable to regard the 22 dog owners who participated in this study as representative of dog owners in general. Do the data from this study provide convincing evidence that there is an increase in mean oxytocin level of dog owners after 30 minutes of interaction with their dogs? State and test the appropriate hypotheses using a significance level of 0.05.

11.27 When the tissue around teeth becomes inflamed, the tissue bleeds more easily and this is an indication of gum disease. The severity of bleeding can be measured by a bleeding score that is determined by the percentage of teeth for which bleeding occurs when the tissue surrounding the tooth is probed. In terms of dental health, lower bleeding scores are better than higher bleeding scores. The authors of the paper **"The Effect of Twice Daily Kiwifruit Consumption on Periodontal and Systemic Conditions Before and After Treatment: A Randomized Clinical Trial" (*Journal of Periodontology* [2018]: 285–293)** wondered if the nutrients found in kiwis might reduce bleeding. To investigate, 25 adult participants agreed to consume 2 kiwis a day for a period of 2 months. A bleeding score was recorded for each participant at the beginning of the study period and again after 2 months of kiwi consumption. Because this resulted in paired samples, differences in bleeding scores (before − after) were analyzed. It was reported that the sample mean difference was $\bar{x}_d = 6.67$ and the standard deviation of the differences was $s_d = 11.90$. Assuming that the 25 study participants can be considered as representative of the population of adults and that the distribution of differences is normal, is there convincing evidence that the mean bleeding score is lower after 2 months of kiwi consumption? Test the relevant hypotheses using a significance level of 0.01.

11.28 Use the information given in the previous exercise to construct and interpret a 90% confidence interval for the difference in bleeding score before and after 2 months of kiwi consumption.

11.29 The paper **"Driving Performance While Using a Mobile Phone: A Simulation Study of Greek Professional Drivers" (*Transportation Research Part F* [2016]: 164–170)** describes a study in which 50 Greek male taxi drivers drove in a driving simulator. In the simulator, they were asked to drive following a lead car. On one drive, they had no distractions and the average distance between the driver's car and the lead car was recorded. In a second drive, the drivers talked on a mobile phone while driving. The authors of the paper used a paired-sampled t test to determine if the mean following distance is greater when the driver has no distractions than when the driver is talking on a mobile phone. The mean of the 50 sample differences (no distraction − talking on mobile phone) was 0.47 meters and the standard deviation of the sample differences was 1.22 meters. The authors concluded that there was evidence to support the claim that the mean following distance for Greek taxi drivers is greater when there are no distractions than when the driver is talking on a mobile phone. Do you agree with this conclusion? Carry out a hypothesis test to support your answer. You may assume that this sample of 50 drivers is representative of Greek taxi drivers.

11.30 The paper referenced in the previous exercise also had the 50 taxi drivers drive in the simulator while sending and receiving text messages. The mean of the 50 sample differences (no distraction − reading text messages) was 1.3 meters and the standard deviation of the sample differences was 1.54 meters. The authors concluded that there was evidence to support the claim that the mean following distance for Greek taxi drivers is greater when there are no distractions than when the driver is texting. Do you agree with this conclusion? Carry out a hypothesis test to support your answer. You can assume that this sample of 50 drivers is representative of Greek taxi drivers.

11.31 Use the information given in the previous exercise to construct and interpret a 95% confidence interval for the difference in mean following distance for Greek taxi drivers while driving with no distractions and while driving and texting.

11.32 The paper **"Quantifying the Benefits of Inefficient Walking: Monty Python Inspired Laboratory Based Experimental Study" (*British Medical Journal* [2022])** carried out a study based on one of Monty Python's most hilarious skits called the Ministry of Silly Walks (take a look at https://www.youtube.com/watch?v=TNeeovY4qNU). In this skit, two silly walks are demonstrated, one by Mr. Putey and one by Mr. Teabag. The authors of this paper wondered if these inefficient (and silly!) walking techniques might actually increase oxygen uptake and energy expenditure compared to casual walking. Thirteen healthy adults participated in the study, and each participant walked for 5 minutes around a track in their usual walking style and then walked for 5 minutes around the track in the Putey style followed by a 5-minute walk in the Teabag style. Oxygen uptake and energy expenditure were measured for each of the three walks using a portable metabolic measurement system.

a. The authors of the paper reported that mean oxygen uptake was significantly higher for the Teabag walk compared to casual walking. Would they have used a paired *t* test or a two-sample *t* test to reach this conclusion?

b. What null hypothesis would they have tested?

c. Did they reject the null hypothesis or fail to reject the null hypothesis?

11.33 ● The humorous paper **"Will Humans Swim Faster or Slower in Syrup?" (*American Institute of Chemical Engineers Journal* [2004]: 2646–2647)** investigates the fluid mechanics of swimming. Twenty swimmers each swam a specified distance in a water-filled pool and in a pool in which the water was thickened with food grade guar gum to create a syrup-like consistency. Velocity, in meters per second, was recorded. Values estimated from a graph that appeared in the paper are given.

The authors of the paper concluded that swimming in guar syrup does not change mean swimming velocity. Are the given data consistent with this conclusion? Carry out a hypothesis test using a 0.01 significance level.

	Velocity (m/s)	
Swimmer	**Water**	**Guar Syrup**
1	0.90	0.92
2	0.92	0.96
3	1.00	0.95
4	1.10	1.13
5	1.20	1.22
6	1.25	1.20
7	1.25	1.26
8	1.30	1.30
9	1.35	1.34
10	1.40	1.41
11	1.40	1.44
12	1.50	1.52
13	1.65	1.58
14	1.70	1.70
15	1.75	1.80
16	1.80	1.76
17	1.80	1.84
18	1.85	1.89
19	1.90	1.88
20	1.95	1.95

11.34 ● The study described in the paper **"Marketing Actions Can Modulate Neural Representation of Experienced Pleasantness" (*Proceedings of the National Academy of Science* [2008]: 1050–1054)** investigated whether price affects people's judgment. Twenty people each tasted six cabernet sauvignon wines and rated how they liked them on a scale of 1 to 6. Prior to tasting each wine, participants were told the price of the wine. Of the six wines tasted, two were actually the same wine, but for one tasting the participant was told that the wine cost $10 per bottle and for the other tasting the participant was told that the wine cost $90 per bottle. The participants were randomly assigned either to taste the $90 wine first and the $10 wine second, or the $10 wine first and the $90 wine second.

Differences were calculated by subtracting the rating for the tasting in which the participant thought the wine cost $10 from the rating for the tasting in which the participant thought the wine cost $90. The differences that follow are consistent with summary quantities given in the paper.

Difference ($90 rating − $10 rating)

2 4 1 2 1 0 0 3 0 2 1 3 3 1 4 1 2 2 1 −1

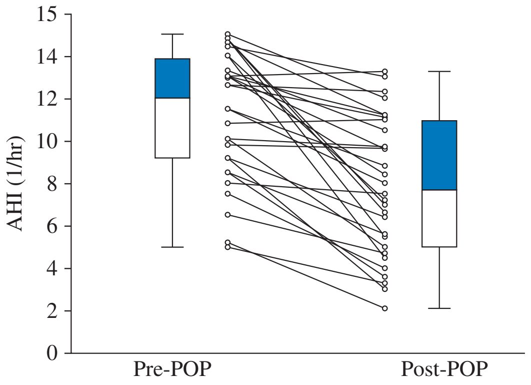
Carry out a hypothesis test to determine if the mean rating assigned to the wine when the cost is described as $90 is greater than the mean rating assigned to the wine when the cost is described as $10. Use $\alpha = 0.01$.

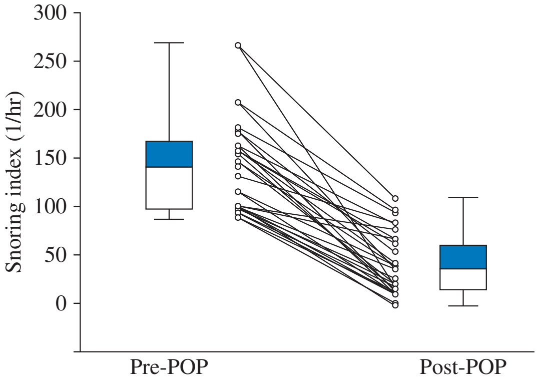
11.35 ● Head movement evaluations are important because disabled individuals may be able to operate communications aids using head motion. The paper **"Constancy of Head Turning Recorded in Healthy Young Humans" (*Journal of Biomedical Engineering* [2008]: 428–436)** reported the accompanying data on neck rotation (in degrees) for 14 subjects both in the clockwise direction (CL) and in the counterclockwise direction (CO). For purposes of this exercise, you can assume that the 14 subjects are representative of the population of adult Americans. Based on these data, is it reasonable to conclude that mean neck rotation is greater in the clockwise direction than in the counterclockwise direction? Carry out a hypothesis test using a significance level of 0.01. (Hint: See Examples 11.6 and 11.7.)

Subject:	1	2	3	4	5	6	7	8	9	10	11	12	13	14
CL:	57.9	35.7	54.5	56.8	51.1	70.8	77.3	51.6	54.7	63.6	59.2	59.2	55.8	38.5
CO:	44.2	52.1	60.2	52.7	47.2	65.6	71.4	48.8	53.1	66.3	59.8	47.5	64.5	34.5

11.36 The paper **"Novel Porous Oral Patched for Patients with Mild Obstructive Sleep Apnea and Mouth Breathing: A Pilot Study" (*Otolaryngology—Head and Neck Surgery* [2015]: 369–373)** describes a study to determine if a porous patch (called a POP) placed over the mouth during sleep would decrease a measure of sleep apnea (AHI) and the intensity of snoring (snoring index, based on the intensity of sound during snoring). The AHI and snoring index were measured for 30 patients sleeping without the patch and sleeping with the patch. The accompanying graphs are similar to ones that appeared in the paper and show the values for AHI and the snoring index both without the patch (pre-POP) and with the patch (post-POP). The pre-POP and post-POP values for each participant are connected by a line. The boxplots that are included summarize the pre-POP measurements and the post-POP measurements, ignoring the pairing.

The authors of the paper concluded that there was a significant decrease in both AHI and the snoring index when the patch was used. Which measure (AHI or snoring index) do you think resulted in the smaller *P*-value? Explain your reasoning.





11.37 ● The authors of the paper **"Statistical Methods for Assessing Agreement Between Two Methods of Clinical Measurement" (*International Journal of Nursing Studies* [2010]: 931–936)** compared two different instruments for measuring a person's ability to breathe out air. (This measurement is helpful in diagnosing various lung disorders.) The two instruments considered were a Wright peak flow meter and a mini-Wright peak flow meter. Seventeen people participated in the study, and for each person air flow was measured once using the Wright meter and once using the mini-Wright meter. The Wright meter is thought to provide a better measure of air flow, but the mini-Wright meter is easier to transport and to use. Use of the mini-Wright meter could be recommended as long as there is not convincing evidence that the mean reading for the mini-Wright meter is different from the mean reading for the Wright meter. Use the given data to determine if there is convincing evidence that the mean reading differs for the two instruments. For purposes of this exercise, you can assume that it is reasonable to consider the 17 Wright meter measurements and the 17 mini-Wright meter measurements as representative samples from their respective populations of measurements.

Subject	Mini-Wright Meter	Wright Meter
1	512	494
2	430	395
3	520	516
4	428	434
5	500	476
6	600	557
7	364	413
8	380	442
9	658	650
10	445	433
11	432	417
12	626	656

(continued)

Subject	Mini-Wright Meter	Wright Meter
13	260	267
14	477	478
15	259	178
16	350	423
17	451	427

11.38 ● The paper **"Quantitative Assessment of Glenohumeral Translation in Baseball Players" (*American Journal of Sports Medicine* [2004]: 1711–1715)** considered various aspects of shoulder motion for a sample of pitchers and another sample of position players. The authors kindly supplied the data shown in the table at the bottom of the page on anteroposterior translation (mm), a measure of the extent of anterior and posterior motion, both for the dominant arm and the nondominant arm.

a. Estimate the true average difference in translation between dominant and nondominant arms for pitchers using a 95% confidence interval.

b. Estimate the true average difference in translation between dominant and nondominant arms for position players using a 95% confidence interval.

c. The authors asserted that pitchers have greater difference in mean anteroposterior translation of their shoulders than do position players. Do you agree? Explain.

11.39 ● Two proposed computer mouse designs were compared by recording wrist extension in degrees for 24 people who each used both mouse types **("Comparative Study of Two Computer Mouse Designs," Cornell Human Factors Laboratory Technical Report RP7992)**. The difference in wrist extension was calculated by subtracting wrist extension for mouse type B from the wrist extension for mouse type A for each person. The mean difference was reported to be 8.82 degrees. Assume that it is reasonable to regard this sample of 24 people as representative of the population of computer users.

a. Suppose that the standard deviation of the differences was 10 degrees. Is there convincing evidence that the mean wrist extension for mouse type A is greater than for mouse type B? Use a 0.05 significance level.

b. Suppose that the standard deviation of the differences was 26 degrees. Is there convincing evidence that the mean wrist extension for mouse type A is greater than for mouse type B? Use a 0.05 significance level.

c. Briefly explain why different conclusions were reached in the hypothesis tests of Parts (a) and (b).

11.40 ● The authors of the paper **"Body Composition Measurement in Young Children Using Quantitative Magnetic Resonance: A Comparison with Air Displacement Plethysmography" (Pediatric Obesity [2018]: 365–373)** compared two different methods for measuring body fat mass. Body fat mass was measured for a representative sample of 5-year-old children using two different methods. One method was

Data for Exercise 11.38

Player	Position Player Dominant Arm	Position Player Nondominant Arm	Pitcher	Pitcher Dominant Arm	Pitcher Nondominant Arm
1	30.31	32.54	1	27.63	24.33
2	44.86	40.95	2	30.57	26.36
3	22.09	23.48	3	32.62	30.62
4	31.26	31.11	4	39.79	33.74
5	28.07	28.75	5	28.50	29.84
6	31.93	29.32	6	26.70	26.71
7	34.68	34.79	7	30.34	26.45
8	29.10	28.87	8	28.69	21.49
9	25.51	27.59	9	31.19	20.82
10	22.49	21.01	10	36.00	21.75
11	28.74	30.31	11	31.58	28.32
12	27.89	27.92	12	32.55	27.22
13	28.48	27.85	13	29.56	28.86
14	25.60	24.95	14	28.64	28.58
15	20.21	21.59	15	28.58	27.15
16	33.77	32.48	16	31.99	29.46
17	32.59	32.48	17	27.16	21.26
18	32.60	31.61			
19	29.30	27.46			

a new method using magnetic resonance (MR) and the other was a commonly used air displacement method (AD). The new method is faster and does not require subjects to remain motionless, making it an attractive alternative for use with children, as long as the measurements are reliable. The data shown below are the difference in body fat mass measurements (in kg, AD − MR) for 12 children (a subset of the data given in a graph that appeared in the paper).

Do these data provide convincing evidence that the mean body fat mass measurement differs for the two methods? Test the relevant hypotheses using $\alpha = 0.05$.

Data for Exercise 13.40

Difference in Body Fat Measurement (AD − MR)	1.3	0.4	−0.6	1.5	−0.3	0.3	−0.3	0.4	−1.1	1.6	0.5	−1.2

Section 11.3 Large-Sample Inferences Concerning the Difference Between Two Population or Treatment Proportions

Large-sample methods for estimating and testing hypotheses about a single population proportion were presented in Chapters 9 and 10. The symbol p was used to represent the proportion of individuals in the population who possess some characteristic (the successes). Inferences about the value of p were based on $\hat{p}$, the corresponding sample proportion of successes.

Many investigations are carried out to compare the proportion of successes in two populations or for two treatments. As was the case for means, the subscripts 1 and 2 are used to distinguish between the two population proportions, sample sizes, and sample proportions.

Notation

Population or Treatment 1: Proportion of "successes" $= p_1$

Population or Treatment 2: Proportion of "successes" $= p_2$

	Sample Size	**Proportion of Successes**
Sample from Population or Treatment 1	n_1	$\hat{p}_1$
Sample from Population or Treatment 2	n_2	$\hat{p}_2$

Comparing two populations or treatments on the basis of "success" proportions focuses on $p_1 - p_2$, the difference between the two proportions. Because $\hat{p}_1$ provides an estimate of p_1 and $\hat{p}_2$ provides an estimate of p_2, the obvious choice for an estimate of $p_1 - p_2$ is $\hat{p}_1 - \hat{p}_2$.

Because $\hat{p}_1$ and $\hat{p}_2$ each vary in value from sample to sample, so does the difference $\hat{p}_1 - \hat{p}_2$. For example, a first sample from each of two populations might yield

$$\hat{p}_1 = 0.69 \qquad \hat{p}_2 = 0.70 \qquad \hat{p}_1 - \hat{p}_2 = 0.01$$

A second sample from each of these populations might result in

$$\hat{p}_1 = 0.79 \qquad \hat{p}_2 = 0.67 \qquad \hat{p}_1 - \hat{p}_2 = 0.12$$

and so on. Because the statistic $\hat{p}_1 - \hat{p}_2$ is the basis for drawing inferences about $p_1 - p_2$, you need to know something about its behavior from one pair of samples to another, which is described by its sampling distribution.

Properties of the Sampling Distribution of $\hat{p}_1 - \hat{p}_2$

For independently selected random samples, the following properties hold:

1. $\mu_{\hat{p}_1-\hat{p}_2} = p_1 - p_2$

   This means that the sampling distribution of $\hat{p}_1 - \hat{p}_2$ is centered at $p_1 - p_2$, so $\hat{p}_1 - \hat{p}_2$ is an unbiased statistic for estimating $p_1 - p_2$.

2. $\sigma^2_{\hat{p}_1-\hat{p}_2} = \sigma^2_{\hat{p}_1} + \sigma^2_{\hat{p}_2} = \dfrac{p_1(1-p_1)}{n_1} + \dfrac{p_2(1-p_2)}{n_2}$

   and

   $$\sigma_{\hat{p}_1-\hat{p}_2} = \sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}}$$

3. If both n_1 and n_2 are large (if $n_1p_1 \geq 10$, $n_1(1-p_1) \geq 10$, $n_2p_2 \geq 10$, and $n_2(1-p_2) \geq 10$), then $\hat{p}_1$ and $\hat{p}_2$ each have a sampling distribution that is approximately normal and the difference $\hat{p}_1 - \hat{p}_2$ also has a sampling distribution that is approximately normal.

The properties in the box imply that when the samples are independently selected and when both sample sizes are large, the statistic

$$z = \frac{(\hat{p}_1 - \hat{p}_2) - (p_1 - p_2)}{\sqrt{\dfrac{p_1(1-p_1)}{n_1} + \dfrac{p_2(1-p_2)}{n_2}}}$$

has a distribution that is approximately standard normal.

A Large-Sample Test Procedure

When comparing p_1 and p_2, the most general null hypothesis of interest has the form

$$H_0\colon\ p_1 - p_2 = \text{hypothesized value}$$

However, when the hypothesized value is something other than 0, the appropriate test statistic differs somewhat from the test statistic used for H_0: $p_1 - p_2 = 0$. Because H_0: $p_1 - p_2 = 0$ is almost always the relevant null hypothesis in applied problems, this is the only case considered here. The test procedures you have seen so far control the probability of a Type I error at the desired level α. This requires using a test statistic with a sampling distribution that is known when H_0 is true. That is, the test statistic should be developed under the assumption that $p_1 = p_2$ (as specified by the null hypothesis $p_1 - p_2 = 0$). If the two population or treatment proportions are equal, they have a common value. This common value is denoted by p. The z statistic given above then simplifies to

$$z = \frac{\hat{p}_1 - \hat{p}_2}{\sqrt{\dfrac{p(1-p)}{n_1} + \dfrac{p(1-p)}{n_2}}}$$

Unfortunately, this z statistic cannot be used as a test statistic, because to calculate the denominator you need to know the value of p. H_0 says that there is a common value p, but it does not specify what that value is. Fortunately, a test statistic can be obtained by first *estimating* p from the sample data and then using this estimate in the denominator of z.

When $p_1 = p_2$, both $\hat{p}_1$ and $\hat{p}_2$ are estimates of the common proportion p. However, a better estimate than either $\hat{p}_1$ or $\hat{p}_2$ is a weighted average of the two, with more weight given to the sample proportion based on the larger sample.

Definition

Combined estimate of a common population proportion:

$$\hat{p}_c = \frac{n_1\hat{p}_1 + n_2\hat{p}_2}{n_1 + n_2} = \frac{\text{total number of } S\text{'s in the two samples}}{\text{total of the two sample sizes}}$$

Copyright 2025 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s). Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

The test statistic for testing $H_0: p_1 - p_2 = 0$ results from using $\hat{p}_c$, the combined estimate, in place of p in the standardized z statistic given previously. This z statistic has approximately a standard normal distribution when H_0 is true, so a P-value can be calculated as an area under the standard normal curve.

Large-Sample z Test for $p_1 - p_2 = 0$

Null Hypothesis: $H_0: p_1 - p_2 = 0$

Test Statistic: $$z = \frac{\hat{p}_1 - \hat{p}_2}{\sqrt{\dfrac{\hat{p}_c(1-\hat{p}_c)}{n_1} + \dfrac{\hat{p}_c(1-\hat{p}_c)}{n_2}}}$$

Alternative Hypothesis:	***P*-Value:**
$H_a: p_1 - p_2 > 0$	Area under the z curve to the right of z
$H_a: p_1 - p_2 < 0$	Area under the z curve to the left of z
$H_a: p_1 - p_2 \neq 0$	2(area to the right of z) if z is positive or 2(area to the left of z) if z is negative

Assumptions:
1. The samples are independently chosen random samples, or treatments were assigned at random.
2. Both sample sizes are large:

$$n_1\hat{p}_1 \geq 10 \quad n_1(1-\hat{p}_1) \geq 10 \quad n_2\hat{p}_2 \geq 10 \quad n_2(1-\hat{p}_2) \geq 10$$

Example 11.9 Electric Cars

The article **"Americans Support Incentives for Electric Vehicles but are Divided Over Buying One Themselves" (Pew Research Center, August 1, 2022)** describes a survey of public opinion on electric vehicles. Each person in a representative sample of adult Americans living in urban areas and each person in an independently selected representative sample of adult Americans living in suburban areas was asked whether they would seriously consider buying an electric vehicle the next time they purchase a vehicle.

In the urban sample, 53% said that they would be likely to consider an electric vehicle. In the suburban sample, 44% responded this way. The article did not give the sample sizes, but for purposes of this example, suppose that the sample sizes were 300.

One question of interest is whether the proportion who would seriously consider buying an electric car is different for people living in urban areas and people living in suburban areas. The following table summarizes what you know so far.

Population	Population Proportion	Sample Size	Sample Proportion
Adult Americans living in urban areas	p_1 = proportion of *all* adult Americans living in urban areas who would seriously consider purchasing an electric car	$n_1 = 300$	$\hat{p}_1 = 0.53$
Adult Americans living in suburban areas	p_2 = proportion of *all* adult Americans living in suburban areas who would seriously consider purchasing an electric car	$n_2 = 300$	$\hat{p}_2 = 0.44$

What Do You Think?

The two sample proportions are not equal. Does this mean that you can automatically conclude that the two population proportions must be different?

Notice that the two sample proportions are not equal. But even if the two population proportions were equal, you wouldn't expect the two sample proportions to be equal because of sampling variability—the differences that occur from one sample to another just by chance. The important question is whether chance is a believable explanation for the observed difference in the two sample proportions, or whether this difference is large enough that you don't think it would have occurred just by chance. A hypothesis test will help you to make this determination.

Understand the context ➤

1. p_1 = proportion of adult Americans living in urban areas who would seriously consider purchasing an electric car

   p_2 = proportion of adult Americans living in suburban areas who would seriously consider purchasing an electric car

   $p_1 - p_2$ is the difference between the proportion who would seriously consider buying an electric car for urban and suburban adult Americans
2. H_0: $p_1 - p_2 = 0$ $(p_1 = p_2)$
3. H_a: $p_1 - p_2 \neq 0$

Formulate a plan ➤

4. Significance level: $\alpha = 0.05$
5. Test statistic:

$$z = \frac{\hat{p}_1 - \hat{p}_2}{\sqrt{\frac{\hat{p}_c(1-\hat{p}_c)}{n_1} + \frac{\hat{p}_c(1-\hat{p}_c)}{n_2}}}$$

6. Assumptions: There are two conditions that need to be met in order for the large sample test for a difference in population proportions to be appropriate.

   The large samples condition is easily verified. The sample sizes are large enough because

$$n_1\hat{p}_1 = 300(0.53) = 159 \qquad n_1(1-\hat{p}_1) = 300(0.47) = 141$$
$$n_2\hat{p}_2 = 300(0.44) = 132 \qquad n_2(1-\hat{p}_2) = 300(0.56) = 168$$

   are all greater than 10.

   From the study description, you know that the samples were independently selected. You also know that the samples were selected in a way that would result in samples that were representative of adult Americans living in urban areas and adult Americans living in suburban areas.

Do the work ➤

7. Calculations: To calculate the value of the test statistic, you need to know the values of the sample proportions and the value of $\hat{p}_c$, the combined estimate of the common population proportion.

$$n_1 = 300 \qquad \hat{p}_1 = 0.53$$
$$n_2 = 300 \qquad \hat{p}_2 = 0.44$$

$$\hat{p}_c = \frac{n_1\hat{p}_1 + n_2\hat{p}_2}{n_1 + n_2} = \frac{300(0.53) + 300(0.44)}{600} = 0.485$$

   You can now calculate the value of the test statistic:

$$\begin{aligned} z &= \frac{\hat{p}_1 - \hat{p}_2}{\sqrt{\frac{\hat{p}_c(1-\hat{p}_c)}{n_1} + \frac{\hat{p}_c(1-\hat{p}_c)}{n_2}}} \\ &= \frac{0.53 - 0.44}{\sqrt{\frac{(0.485)(0.515)}{300} + \frac{(0.485)(0.515)}{300}}} \\ &= \frac{0.09}{0.041} = 2.20 \end{aligned}$$

8. *P*-value: This is a two-tailed test (the inequality in H_a is $\neq$), so the *P*-value is twice the area under the *z* curve and to the right of the calculated *z* value:

$$\begin{aligned} P\text{-value} &= 2(\text{area under } z \text{ curve to the right of } 2.20) \\ &= 2 \cdot P(z > 2.20) \end{aligned}$$

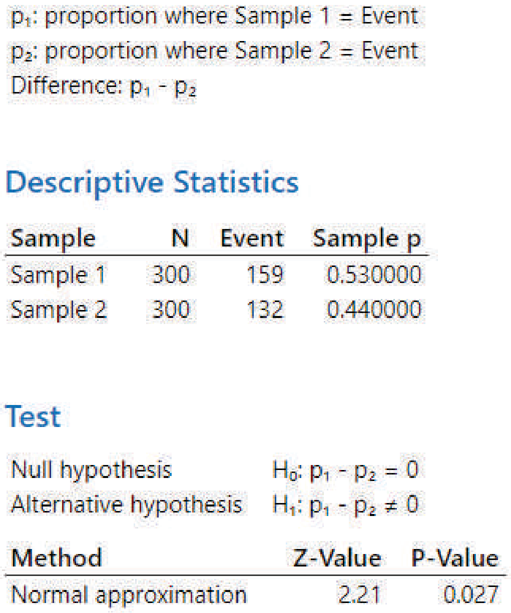
Copyright 2025 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s). Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

$$= 2(0.0139)$$
$$= 0.0278$$

Interpret the results ➤

9. Because the P-value (0.0278) is less than the selected significance level (0.05), you reject the null hypothesis. There is convincing evidence to conclude that the proportion of adult Americans living in urban areas who would seriously consider buying an electric car is not the same as this proportion for adult Americans living in suburban areas.

It is also possible to use statistical software or a graphing calculator to calculate the value of the test statistic and the P-value. For example, Minitab output is shown here.



```
p₁: proportion where Sample 1 = Event
p₂: proportion where Sample 2 = Event
Difference: p₁ - p₂


Descriptive Statistics

Sample      N    Event   Sample p
Sample 1   300    159    0.530000
Sample 2   300    132    0.440000


Test

Null hypothesis          H₀: p₁ - p₂ = 0
Alternative hypothesis   H₁: p₁ - p₂ ≠ 0

Method                  Z-Value   P-Value
Normal approximation       2.21     0.027
```

From the Minitab output, $z = 2.21$ and the associated P-value is 0.027. These values are slightly different from those calculated by hand only because Minitab uses greater decimal accuracy in the calculations leading to the value of the test statistic.

Example 11.10 Debt and Dating

Would you reconsider a romantic relationship if you found out that the person you were dating had a lot of debt? The folks at Finder, an online financial information service, conducted a survey of representative samples of adult Americans to find out how debt might be a factor in the decision to begin a new relationship or to rethink a current relationship. The article **"Is Your Debt Stopping You From Finding Love?" (finder.com, April 26, 2021)** states that the proportion of those in a sample of millennials (defined in the article as those born between 1981 and 1996) who said that they would reconsider a romantic relationship due to a partner's debt was 0.342 (or 34.2%). The proportion was 0.257 (or 25.7%) for a sample of Generation Z (those born between 1997 and 2003).

The article does not say how large the sample sizes were, but the description of the survey methodology indicates that the samples can be regarded as independent random samples. The article does give the total sample size for all adults surveyed (which also included Generation X, Baby Boomers, and the Silent Generation) as 1658. For purpose of this example, suppose that both sample sizes were 300.

One question of interest is whether there is convincing evidence that the proportion who would reconsider a romantic relationship because of a partner's debt is greater for the millennial population than for the Generation Z population. To answer this question, you can carry out a hypothesis test. For this example, a significance level of 0.05 will be used.

Understand the context ➤

1. p_1 = proportion of all millennials who would reconsider relationship
p_2 = proportion all Generation Z who would reconsider relationship

2. H_0: $p_1 - p_2 = 0$
3. H_a: $p_1 - p_2 > 0$
4. Significance level: $\alpha = 0.05$

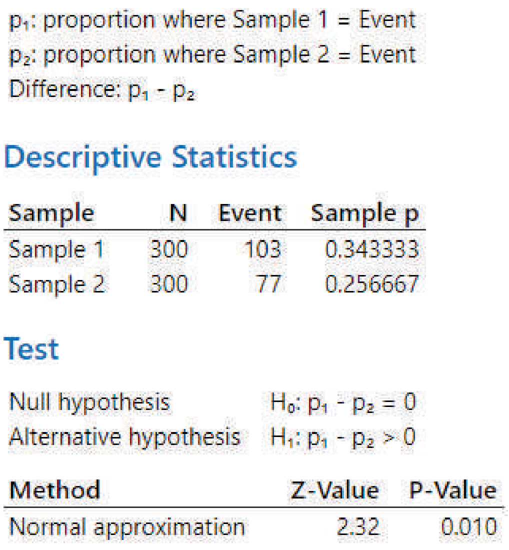
Formulate a plan ➤

5. Test statistic: $z = \dfrac{\hat{p}_1 - \hat{p}_2}{\sqrt{\dfrac{\hat{p}_c(1-\hat{p}_c)}{n_1} + \dfrac{\hat{p}_c(1-\hat{p}_c)}{n_2}}}$
6. Assumptions: The two samples can be regarded as independent random samples. Checking to make sure that the sample sizes are large enough by using $n_1 = 300$, $\hat{p}_1 = 0.342$, $n_2 = 300$, and $\hat{p}_2 = 0.257$ results in

$$\begin{aligned} n_1\hat{p}_1 &= 102.6 \geq 10 \\ n_1(1-\hat{p}_1) &= 197.4 \geq 10 \\ n_2\hat{p}_2 &= 77.1 \quad \geq 10 \\ n_2(1-\hat{p}_2) &= 222.9 \geq 10 \end{aligned}$$

Do the work ➤

7. Calculations: Minitab output is shown below. From the output, $z = 2.32$.



```
p₁: proportion where Sample 1 = Event
p₂: proportion where Sample 2 = Event
Difference: p₁ - p₂

Descriptive Statistics

Sample      N    Event   Sample p
Sample 1   300    103    0.343333
Sample 2   300     77    0.256667

Test

Null hypothesis          H₀: p₁ - p₂ = 0
Alternative hypothesis   H₁: p₁ - p₂ > 0

Method                  Z-Value   P-Value
Normal approximation       2.32     0.010
```

8. P-value: From the computer output, P-value $= 0.010$

Interpret the results ➤

9. Conclusion: Because the P-value (0.010) is less than α (0.05), the null hypothesis is rejected. There is convincing evidence that the proportion who would reconsider a romantic relationship due to a partner's debt is greater for millennials than for Generation z.

A Confidence Interval

A large-sample confidence interval for $p_1 - p_2$ is a special case of the general z interval formula

point estimate ± (z critical value)(estimated standard deviation)

The statistic $\hat{p}_1 - \hat{p}_2$ gives a point estimate of $p_1 - p_2$, and the standard deviation of this statistic is

$$\sigma_{\hat{p}_1 - \hat{p}_2} = \sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}}$$

An estimated standard deviation is obtained by using the sample proportions $\hat{p}_1$ and $\hat{p}_2$ in place of p_1 and p_2 under the square-root symbol. Notice that this estimated standard deviation differs from the one used previously in the test statistic. When constructing a

Copyright 2025 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s). Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

confidence interval, there isn't a null hypothesis that claims $p_1 = p_2$, so there is no assumed common value of p to estimate.

Large-Sample Confidence Interval for $p_1 - p_2$

When

1. the samples are independently selected random samples or treatments were assigned at random to individuals or objects (or vice versa), and
2. both sample sizes are large:

$$n_1\hat{p}_1 \geq 10 \qquad n_1(1-\hat{p}_1) \geq 10 \qquad n_2\hat{p}_2 \geq 10 \qquad n_2(1-\hat{p}_2) \geq 10$$

a large-sample confidence interval for $p_1 - p_2$ is

$$(\hat{p}_1 - \hat{p}_2) \ \pm \ (z \text{ critical value})\sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}}$$

Example 11.11 Concern about Climate Change

Understand the context ➤

The report **"U.S. Concern about Climate Change Is Rising, But Mainly among Democrats" (Pew Research Center, April 16, 2020)** described a survey of adults living in the United States. Each person in a representative sample of 487 Democrats and a representative sample of 429 Republicans was asked if they viewed climate change as a major threat to the well-being of the United States. It was reported that 428 (88%) of the Democrats surveyed and 133 (31%) of the Republicans surveyed said that they viewed climate change as a major threat. The two samples were independently selected and were thought to be representative of Democrats and Republicans. This information can be used to estimate the difference between the proportion of Democrats who view climate change as a major threat, p_1, and this proportion for Republicans, p_2.

Formulate a plan ➤

To estimate this difference, you can construct a 95% confidence interval for a difference in population proportions.

There are two conditions that need to be met in order to use the large sample confidence interval for a difference in population proportions. The large sample condition is easily verified. With "success" denoting a person who views climate change as a major threat, the sample sizes are large enough because there are more than 10 successes (428 in sample 1 and 133 in sample 2) and more than 10 failures (487 − 428 = 59 in sample 1 and 429 − 133 = 296 in sample 2) in each sample. This is equivalent to checking that $n_1\hat{p}_{1,}$ $n_1(1-\hat{p}_1)$, $n_2\hat{p}_2$, and $n_2(1-\hat{p}_2)$ are all greater than 10.

The study description stated that the samples are representative of the two populations of interest and were independently selected.

Do the work ➤

A 95% confidence interval for $p_1 - p_2$ can now be calculated:

$$n_1 = 487 \qquad\qquad n_2 = 429$$

$$\hat{p}_1 = \frac{428}{487} = 0.88 \qquad \hat{p}_2 = \frac{133}{429} = 0.31$$

For a confidence level of 95% the appropriate z critical value is 1.96. Substituting the values for n_1, n_2, $\hat{p}_1$, and $\hat{p}_2$ into the 95% confidence interval formula results in

$$(\hat{p}_1 - \hat{p}_2) \pm (z \text{ critical value})\sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}}$$

$$(0.88 - 0.31) \pm (1.96)\sqrt{\frac{(0.88)(1-0.88)}{487} + \frac{(0.31)(1-0.31)}{429}}$$

$$0.57 \pm (1.96)\sqrt{0.00072}$$

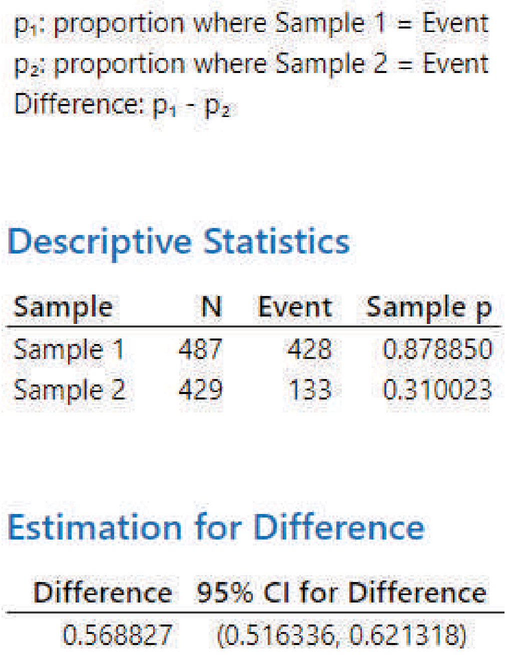
$$0.57 \pm (1.96)(0.027)$$

$$0.57 \pm 0.053$$

$$(0.517, 0.623)$$

Copyright 2025 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s). Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

Statistical software or a graphing calculator could also have been used to compute the end points of the confidence interval. Minitab output is shown here. (Minitab has used more decimal accuracy in computing the endpoints of the confidence interval.)



```
p₁: proportion where Sample 1 = Event
p₂: proportion where Sample 2 = Event
Difference: p₁ - p₂


Descriptive Statistics

Sample      N    Event   Sample p
Sample 1   487    428    0.878850
Sample 2   429    133    0.310023


Estimation for Difference

 Difference   95% CI for Difference
   0.568827   (0.516336, 0.621318)
```

Interpret the results ➤

You can be 95% confident that the actual difference in the proportion of Democrats and the proportion of Republicans who view climate change as a major threat to the well-being of the United States is between 0.516 and 0.621. Since both endpoints of the confidence interval are positive, this means that you think that the proportion of Democrats who view climate change as a major threat is greater than this proportion for Republicans by somewhere between 0.516 and 0.621. This is a very large difference. The method used to construct this estimate is successful in capturing the actual difference in population proportions about 95% of the time.

Exercises 11.41–11.59

11.41 Some people believe that you can fix anything with duct tape. Even so, many are skeptical when they are told that duct tape may be a more effective and less painful alternative to freezing with liquid nitrogen, which doctors routinely use to treat warts. The paper **"Comparison of Duct Tape with Adhesive Cyanoacrylate Versus Cryotherapy in the Treatment of Palmoplantar Warts" (*Journal of Advances in Medical and Pharmaceutical Sciences,* [2016]: 1–6)** describes a study in which patients with warts were randomly assigned to one of two groups. The patients in the duct tape group applied a film of cyanoacrylate (which is superglue!) to the wart and then covered it with duct tape. The tape was left in place for 6 days, and at the end of 6 days the tape was removed, the area was soaked in warm water, and then rubbed with a pumice stone or a nail file. This process was repeated for up to 2 months or until the wart was gone, whichever occurred first. The patients in the freezing group had liquid nitrogen applied to the wart for 10 to 20 seconds every 2 to 3 weeks for up to 2 months or until the wart was gone. Data consistent with summary values given in the paper are given in the following table:

Treatment	***n***	**Number with Wart Successfully Removed**
Liquid nitrogen freezing	100	60
Duct tape	100	80

Do these data suggest that freezing is less successful than duct tape in removing warts? Test the relevant hypotheses using a significance level of 0.01. (Hint: See Example 11.9.)

11.42 The paper **"Tezepelumab Improves Patient-Reported Outcomes in Patients with Severe, Uncontrolled Asthma in PATHWAY" (*Annals of***

Copyright 2025 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s). Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

Allergy Asthma Immunology [2021]: 187–193) describes an experiment to evaluate the effectiveness of the drug Tezepelumab for patients with severe asthma. Patients with severe uncontrolled asthma were randomly assigned to groups. One group of 412 patients received the drug and 138 patients in a second group received a placebo (a treatment with no active ingredients). After 1 year, all patients were classified based on whether their asthma was well-controlled, partially controlled, or uncontrolled. The paper states that "The proportions of patients with well-controlled or partially-controlled asthma were higher in the Tezepelumab-treated group than in the placebo group." The difference in the proportions with well-controlled asthma for the group that received the drug treatment (0.233) and the group that received the placebo treatment (0.138) was found to be statistically significant. Once this is determined, a logical next step is to estimate the difference in the proportions for these two treatments (drug and placebo) using a confidence interval. Estimate this difference using a 90% confidence interval.

11.43 The report **"The 2016 Inside Higher Ed Survey of Faculty Attitudes on Technology" (Inside Higher Ed, 2016)** describes the results of a survey of 1129 full-time college faculty and 293 part-time college faculty. Survey participants were asked if they require undergraduate students to submit papers through plagiarism-detection software; 40% of the full-time faculty and 38% of the part-time faculty said "yes." Is there convincing evidence that the proportion who require students to submit papers through plagiarism-detection software is different for full-time faculty and part-time faculty? You may assume that the samples were independently selected and that they are representative of the two populations of faculty. Use a significance level of 0.05.

11.44 The article **"Country Living Enjoys Renewed Appeal in U.S." (gallup.com, January 5, 2021)** describes results from a representative sample of American adults in 2018 and an independently selected representative sample of American adults taken in 2019. Survey participants were asked the following question:

> If you could live anywhere you wished, where would you prefer to live—in a big city, small city, suburb of a big city, suburb of a small city, or in a town or rural area?

In 2018, 390 of the 1000 people surveyed selected town or rural area. In 2019, 480 of the 1000 people surveyed selected town or rural area. Use a significance level of 0.01 to determine if there is convincing evidence that the proportion of all American adults who prefer to live in a rural area increased from 2018 to 2019.

11.45 The article referenced in the previous exercise also reported that 52% of the people surveyed in 2019 said they would prefer to live in a city or a suburb of a city. Would you use the large-sample test for a difference in population proportions to test the hypothesis that a majority of all American adults prefers to live in a city or a suburb? Explain why or why not.

11.46 The authors of the paper **"Appointment Reminders to Increase Uptake of HIV Retesting by At-Risk Individuals" (*Journal of the International AIDS Society* [2020])** carried out an experiment to determine if appointment reminders were an effective way to increase the proportion of patients testing for HIV who would show up for follow-up testing. At-risk patients who were told that they should retest at a later date were randomly assigned to one of three groups. One group was encouraged to return for retesting, but no follow-up visit was scheduled and no reminder to schedule was sent. A second group was encouraged to return and they received a reminder to schedule an appointment 4 months later. Those in the third group were given a return appointment for testing and were sent a reminder prior to the appointment. Of the 215 people in the no appointment, no reminder group, 24 had returned for retesting within 7 months. Of the 218 people in the group who did not have a scheduled appointment but who received a reminder to schedule a return appointment, 42 had returned within 7 months. Do these data support the hypothesis that receiving a reminder is effective in increasing the proportion who return for retesting?

11.47 According to the **U.S. Census Bureau (www.census.gov)**, the percentage of U.S. residents living in poverty in 2020 was 16.1% for people under age 18 and 10.4% for people of age 18 to 64. These percentages were estimates based on data from large representative samples. Suppose that the sample sizes were 1200 for people under age 18 and 1000 for people of age 18 to 64. You would like to use the survey data to estimate the difference in the proportion living in poverty in 2020 for these two groups. Calculate and interpret a 90% large-sample confidence interval for the difference in the proportion living in poverty in 2020 for people under age 18 and people of age 18 to 64.

11.48 The **U.S. Department of Transportation** reported that in studies of mobile phone use while driving conducted in 2015 and 2020, 3.8% of the drivers observed in 2015 and 2.6% of the drivers observed in 2020 were using a handheld mobile phone while driving **("Driver Electronic Device Use in 2020," November 2021)**.

Suppose that these percentages were based on observations from independent random samples of 1200 drivers in 2015 and 1200 drivers in 2020 (the actual sample sizes were much larger).

a. Are the sample sizes large enough to use the large-sample confidence interval for a difference in population proportions?

b. Assume that it is reasonable to regard these samples as representative of drivers in 2015 and drivers in 2020. Estimate the difference in the proportion of drivers in 2015 and the proportion of drivers in 2020 who use a mobile phone while driving using a 95% confidence interval.

11.49 The paper **"App-Based Relaxation Exercises for Patients With Chronic Neck Pain: Pragmatic Randomized Trial" (*JMIR mHealth and uHealth* [2021])** describes an experiment to investigate whether app-based relaxation exercises are effective in reducing neck pain. Participants in the experiment were 220 people with chronic neck pain and participants were randomly assigned to one of two groups. People in the treatment group received usual care and were also instructed to practice app-based relaxation exercises, while people in the control group received usual care, but were not given the app-based relaxation exercises. The paper states that after 3 months, 60% of the participants in the treatment group reported that they felt their neck pain had improved and 30% of the participants in the control group reported improvement.

a. Suppose that there were 110 people in each of the two groups. Use the given information to construct and interpret a 95% confidence interval for the difference in the proportion of people who would report that their neck pain improved for those who are given the app-based relaxation exercises and those who are not.

b. Does the interval from Part (a) support the conclusion that people using the app-based relaxation exercises are more likely to report improvement? Explain how you used the confidence interval to answer this question.

11.50 The article **"Some Digital Divides Persist Between Rural, Urban and Suburban America" (pewresearch.org, August 19, 2021)** describes surveys of representative samples of rural and urban adult Americans. Of 525 urban adults, 194 reported that they use the Internet almost constantly. In the sample of 280 rural adults, 64 reported that they use the Internet almost constantly. Construct and interpret a 99% confidence interval for the difference in the proportion who report using the Internet almost constantly for urban and rural adult Americans.

11.51 The report **"Global Trustworthiness Index 2021" (ipsos.com, October 12, 2021)** summarizes data from independent representative samples of approximately 1000 people in each of 28 different countries. People in the samples were asked if they thought that, in general, scientists in their country were trustworthy or untrustworthy. In the sample from the United States, 62% responded trustworthy. In the sample from Canada, 64% responded trustworthy. For this exercise, assume that the two sample sizes were 1000. Carry out a hypothesis test to determine if there is convincing evidence that the proportion of people who think that in general, scientists in their country are trustworthy differs for the United States and Canada.

11.52 The report referenced in the previous exercise also asked people if they thought that, in general, teachers in their country were trustworthy or untrustworthy. In the sample from the United States, 57% responded trustworthy. In the sample from Canada, 62% responded trustworthy. Based on these data, can you conclude that the proportion of people in the United States who think that teachers in their country are trustworthy is less than this proportion for people in Canada?

11.53 The report **"The New Food Fights: U.S. Public Divides Over Food Science" (December 1, 2016, pewinternet.org)** states that younger adults are more likely to see foods with genetically modified ingredients as being bad for their health than older adults. This statement is based on a representative sample of 178 adult Americans age 18 to 29 and a representative sample of 427 adult Americans age 50 to 64. Of those in the 18 to 29 age group, 48% said they believed these foods were bad for their health, while only 38% of those in the 50 to 64 age group believed this.

a. Are the sample sizes large enough to use the large-sample confidence interval to estimate the difference in the population proportions?

b. Estimate the difference in the proportion of adult Americans age 18 to 29 who believe the foods made with genetically modified ingredients are bad for their health and the corresponding proportion for adult Americans age 50 to 64. Use a 90% confidence interval.

c. Is zero in the confidence interval? What does this suggest about the difference in the two population proportions?

11.54 A survey of high school students is described in the report **"Students on STEM" (changetheequation.org)**. The report states that 14% of those in a sample of students in low-income households (defined as a household income less than $50,000 per year) and 24% of those in a sample of students in higher income households (defined as a household income of $50,000 or more) participated in a science club or group. Suppose that these samples are representative of high school students in the two income groups and that the two sample sizes were both 500. Use a 95% confidence interval to estimate the difference in the proportion participating in a science club for students in the two income groups.

11.55 *Choice blindness* is the term that psychologists use to describe a situation in which a person expresses a preference and then doesn't notice when they receive something different than what they asked for. The authors of the paper **"Can Chocolate Cure Blindness? Investigating the Effect of Preference Strength and Incentives on the Incidence of Choice Blindness" (*Journal of Behavioral and Experimental Economics* [2016]: 1–11)** wondered if choice blindness would occur more often if people made their initial selection by looking at pictures of different kinds of chocolate compared with if they made their initial selection by looking at the actual different chocolate candies.

Suppose that 200 people were randomly assigned to one of two groups. The 100 people in the first group are shown a picture of eight different kinds of chocolate candy and asked which one they would like to have. After they selected, the picture is removed and they are given a chocolate candy, but not the one they actually selected. The 100 people in the second group are shown a tray with the eight different kinds of candy and asked which one they would like to receive. Then the tray is removed and they are given a chocolate candy, but not the one they selected.

If 20 of the people in the picture group and 12 of the people in the actual candy group failed to detect the switch, would you conclude that there is convincing evidence that the proportion who experience choice blindness is different for the two treatments (choice based on a picture and choice based on seeing the actual candy)? Test the relevant hypotheses using a 0.01 significance level.

11.56 **The Bureau of Labor Statistics ("Education Pays," 2020)** reported that in 2020, 5.5% of people with a bachelor's degree and 9.0% of people with a high school diploma but no college were unemployed. Suppose that the reported percentages were based on independently selected representative samples of 500 people with a bachelor's degree and 500 people with a high school degree but no college. Construct and interpret a 95% confidence interval for the difference in the proportion who were unemployed for these two groups.

11.57 The Bureau of Labor Statistics report referenced in the previous exercise also reported that 4.1% of those with master's degrees were unemployed in 2020. Suppose that the reported percentage was based on a representative sample of 300 people with master's degrees. Use this information and the information provided in the previous exercise to answer the following questions.

a. Construct and interpret a 99% large-sample confidence interval for the difference in the proportion who were unemployed in 2020 for those who have a bachelor's degree and those who have a master's degree.

b. Is the confidence interval from Part (a) wider or narrower than the confidence interval calculated in the previous exercise? Give two reasons why it is wider or narrower.

11.58 People diagnosed with breast cancer whose tumors have not spread may be faced with a decision between two surgical treatments—mastectomy (removal of the breast) or lumpectomy (only the tumor is removed). In a long-term study of the effectiveness of these two treatments, 701 people with breast cancer were randomly assigned to one of two treatment groups. One group received mastectomies and the other group received lumpectomies and radiation. Both groups were followed for 20 years after surgery. It was reported that there was no statistically significant difference in the proportion surviving for 20 years for the two treatments **(*Associated Press*, October 17, 2002)**.

What hypotheses do you think the researchers tested in order to reach the given conclusion? Did the researchers reject or fail to reject the null hypothesis?

11.59 Gallup surveyed adult Americans about their consumer debt **("Americans' Big Debt Burden Growing, Not Evenly Distributed," February 4, 2016, gallup.com)**. The article reported that 47% of millennials (those born between 1980 and 1996) and 61% of Gen Xers (those born between 1965 and 1971) did not pay off their credit cards each month, and therefore carried a balance from month to month. Suppose that these percentages were based on representative samples of 450 millennials and 300 Gen Xers. Is there convincing evidence that the proportion of Gen Xers who do not pay off their credit cards each month is greater than this proportion for millennials? Test the appropriate hypotheses using a significance level of 0.05.

Section 11.4 Interpreting and Communicating the Results of Statistical Analyses

Many different types of research involve comparing two populations or treatments. It is easy to find examples of the two-sample hypothesis tests introduced in this chapter in published sources in a wide variety of disciplines.

Communicating the Results of Statistical Analyses

As was the case with one-sample hypothesis tests, it is important to include a description of the hypotheses, the test procedure used, the value of the test statistic, the *P*-value, and a conclusion in context when summarizing the results of a two-sample test.

Correctly interpreting confidence intervals in the two-sample case is more difficult than in the one-sample case, so take particular care when providing a two-sample confidence interval interpretation. Because the two-sample confidence intervals of this chapter estimate a difference ($\mu_1 - \mu_2$ or $p_1 - p_2$), the most important thing to note is whether or not the interval includes 0. If both endpoints of the interval are positive, then it is correct to say that, based on the interval, you believe that μ_1 is greater than μ_2 (or that p_1 is greater than p_2 if you are working with proportions) and then the interval provides an estimate of how much greater. Similarly, if both interval endpoints are negative, you would say that μ_1 is less than μ_2 (or that p_1 is less than p_2), with the interval providing an estimate of the size of the difference. If 0 is included in the interval, it is plausible that μ_1 and μ_2 (or p_1 and p_2) are equal.

Interpreting the Results of Statistical Analyses

As with one-sample tests, it is common to find only the value of the test statistic and the associated *P*-value (or sometimes only the *P*-value) in published reports. You may have to think carefully about the missing steps to determine whether or not the conclusions are justified.

What to Look for in Published Data

Here are some questions to consider when you are reading a report that contains the result of a two-sample hypothesis test or confidence interval:

- Are only two groups being compared? If more than two groups are being compared two at a time, then a different type of analysis is preferable (see Chapter 15).
- Were the samples selected independently, or were the samples paired? If the samples were paired, was the analysis that was performed appropriate for paired samples?
- If a confidence interval is reported, is it correctly interpreted as an estimate of a difference in population or treatment means or proportions?
- What hypotheses are being tested? Is the test one- or two-tailed?
- Does the validity of the test performed depend on any assumptions about the sampled populations (such as normality)? If so, do the assumptions appear to be reasonable?
- What is the *P*-value associated with the test? Does the *P*-value lead to rejection of the null hypothesis?
- Are the conclusions consistent with the results of the hypothesis test? In particular, if H_0 was rejected, does this indicate practical significance or only statistical significance?

For example, the article **"AI More Accurate at Assessing Heart Health Than Humans, Study Reveals" (studyfinds.org, April 7, 2023)** summarizes results from a study conducted by researchers at Cedars Sinai Medical Center. These researchers conducted an experiment to investigate whether artificial intelligence (AI) is better at interpreting echocardiograms than human ultrasound technicians. Cardiologists reviewed the initial assessment made by either AI or an ultrasound technician and then confirmed the initial assessment of heart health or corrected the initial assessment. The summary in the StudyFinds article was based on the paper **"Blinded, Randomized Trial of Sonographer versus AI Cardiac Function Assessment" (*Nature*, April 5, 2023)**. The *Nature* paper includes the following statement:

> From 3,769 echocardiographic studies screened, 274 studies were excluded owing to poor image quality. The proportion of studies substantially changed was 16.8% in the AI group and 27.2% in the sonographer group (difference of −10.4%, 95% confidence interval: −13.2% to −7.7%.

The paper also indicates that the assessments were randomly assigned to cardiologists for review and that the cardiologists did not know whether they were reviewing an assessment made by AI or an ultrasound technician.

Because the treatments (initial assessment by AI or by ultrasound technician) were randomly assigned to the reviewing cardiologists and the sample sizes were large (1740 for the AI group and 1755 for the ultrasound technician group), use of the two-sample z confidence interval was appropriate. The confidence interval reported in the paper was given in terms of percentages, but you can easily convert from percentages to proportions. Both endpoints of the reported confidence interval were negative $(-0.132, -0.077)$, so you can interpret this interval as meaning that the proportion needing correction for AI initial assessment is less than the proportion needing correction for the ultrasound technician initial assessment by somewhere between 0.077 and 0.132, or equivalently between 7.7 and 13.2 percentage points. That the study included blinding (the reviewing cardiologists did not know if the initial assessment was made by AI or an ultrasound technician) is also a strength of this study.

The StudyFinds summary of the research makes the following statement: "AI was 10% more accurate than human technicians." Unfortunately, this is a misleading statement and makes a common error when interpreting percentages and proportions. There is a difference between a 10 percentage point difference and a 10% difference. A 10 percentage point decrease is indicating that the actual amount of the decrease (as was the case here where 16.8% is about 10 percentage points less than 27.2%). A 10% decrease indicates that the value for AI would be what you get when you reduce the value for ultrasound technicians by 10% *of that value*. So, if that were the case, 10% of 27.2 is 2.72, and the actual difference would be only 2.72 percentage points.

As another example, consider the study described in the paper **"The Effect of Self-Monitoring Limited Social Media Use on Psychological Well-Being" (*Technology, Mind, and Behavior* [2023])**. To investigate the effect of self-monitoring and limiting time spent on social media, researchers randomly assigned students who agreed to participate in the study to one of two groups. One group (the treatment group) was to self-monitor social media use and try to limit use to 30 minutes per day for a 2-week period. The other group (the control group) did not make any changes to their social media use. The sample sizes for the two groups were reported as 99 for the treatment group and 131 for the control group. Several measures of well-being were recorded at the start of the 2 weeks and again at the end of the 2-week period. On a measure of anxiety, the researchers reported that there was no significant difference in mean anxiety score (control – treatment) at the beginning of the study and reported a test statistic value of $t = 0.77$. However, at the end of the 2-week period, the mean anxiety score for the self-monitoring group was significantly lower than the mean for the control group ($t = 4.2$, P-value < 0.001).

Participants were all willing to participate in the experiment and agreed to self-monitor and limit if assigned to the treatment group. Because there was random assignment to the two experimental groups and the sample sizes were large, use of the two-sample t test was appropriate. No P-value accompanied the test statistic value of $t = 0.77$, but you could use technology to calculate it and you would find the P-value ≈ 0.44, which is consistent with a decision not to reject the null hypothesis of no difference in mean anxiety score prior to the start of self-monitoring. The very small P-value for the test at the end of the self-monitoring period would lead to rejecting a null hypothesis of no difference in mean anxiety score for the two groups and is consistent with the conclusion that the mean anxiety score was significantly lower for the self-monitoring group after 2 weeks of self-monitoring and limiting social media use.

In a newspaper article, you may find even less information than in a journal article. For example, the article **"Prayer Is Little Help to Some Heart Patients, Study Shows" (*Chicago Tribune*, March 31, 2006)** included the following paragraphs:

> Bypass patients who consented to take part in the experiment were divided randomly into three groups. Some patients received prayers but were not informed of that. In the second group the patients got no prayers, and also were not informed one way or the other. The third group got prayers and were told so.
>
> There was virtually no difference in complication rates between the patients in the first two groups. But the third group, in which patients knew they were receiving prayers, had a complication rate of 59 percent—significantly more than the rate of 52 percent in the no-prayer group.

Earlier in the article, the total number of participants in the experiment was given as 1800. The author of this article has done a good job of describing the important aspects of the experiment. The final comparison in the quoted paragraph was probably based on a two-sample z test for proportions, comparing the sample proportion with complications for the 600 patients in the no-prayer group with the sample proportion with complications for the 600 participants who knew that someone was praying for them.

For the reported sample sizes and sample proportions, the test statistic for testing H_0: $p_1 - p_2 = 0$ versus H_a: $p_1 - p_2 < 0$ (where p_1 represents the complication proportion for patients who did not receive prayers and p_2 represents the complication proportion for patients who knew they were receiving prayers) is $z = -2.10$. The associated P-value is 0.036, supporting the conclusion stated in the article.

A Word to the Wise: Cautions and Limitations

The three cautions that appeared at the end of Chapter 10 apply here as well. They were (see Chapter 10 for more details):

1. Remember that the result of a hypothesis test can never show strong support for the null hypothesis. In two-sample situations, this means that you shouldn't be *convinced* that there is no difference between population means or proportions based on the outcome of a hypothesis test.
2. If you have complete information (a census) of both populations, there is no need to carry out a hypothesis test or to construct a confidence interval—in fact, it would be inappropriate to do so.
3. Don't confuse statistical significance and practical significance. In the two-sample setting, it is possible to be convinced that two population means or proportions are not equal even in situations where the actual difference between them is small enough that it is of no practical interest. After rejecting a null hypothesis of no difference (statistical significance), it is useful to look at a confidence interval estimate of the difference to get a sense of practical significance.

And here's one new caution to keep in mind for two-sample tests:

4. Be sure to think carefully about how the data were collected, and make sure that an appropriate test procedure or confidence interval is used. A common mistake is to overlook pairing and to analyze paired samples as if they were independent. It is usually easy to tell if the samples are paired—you just have to remember to think about how the samples were selected.

Exercises 11.60–11.62

11.60 The paper **"The Psychological Consequences of Money" (*Science* [2006]: 1154–1156)** describes several experiments designed to investigate the way in which money can change behavior. In one experiment, participants completed one of two versions of a task in which they were given lists of five words and were asked to rearrange four of the words to create a sensible phrase. For one group, half of the 30 unscrambled phrases related to money, whereas the other half were phrases that were unrelated to money. For the second group (the control group), none of the 30 unscrambled phrases related to money. Participants were 44 students at Florida State University. Participants

Copyright 2025 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s). Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

received course credit and \$2 for their participation. The following description of the experiment is from the paper:

> Participants were randomly assigned to one of two conditions, in which they descrambled phrases that primed money or neutral concepts. Then participants completed some filler questionnaires, after which the experimenter told them that the experiment was finished and gave them a false debriefing. This step was done so that participants would not connect the donation opportunity to the experiment. As the experimenter exited the room, she mentioned that the lab was taking donations for the University Student Fund and that there was a box by the door if the participant wished to donate. Amount of money donated was the measure of helping. We found that participants primed with money donated significantly less money to the student fund than participants not primed with money [$t(38) = 2.13$, $P < 0.05$].

The paper also gave the following information on amount donated for the two experimental groups.

Group	Mean	Standard Deviation
Money primed	\$0.77	\$0.74
Control	\$1.34	\$1.02

- **a.** Explain why the random assignment of participants to experimental groups is important in this experiment.
- **b.** Use the given information to verify the values of the test statistic and degrees of freedom (38, given in parentheses just after the t in the quote from the paper) and the statement about the P-value. Assume that both sample sizes are 22.
- **c.** Do you think that use of the two-sample t test was appropriate in this situation? *Hint*: Are the assumptions required for the two-sample t test reasonable?

11.61 An experiment to determine if an online intervention can reduce references to sex and substance abuse on social networking websites of adolescents is described in the paper **"Reducing At-Risk Adolescents' Display of Risk Behavior on a Social Networking Website" (*Archives of Pediatrics and Adolescent Medicine* [2009]: 35–41)**. Researchers selected public profiles of people who described themselves as between 18 and 20 years old and who referenced sex or substance use (alcohol or drugs) in their profiles. The selected subjects were assigned at random to an intervention group or a control group.

Those in the intervention group were sent an e-mail from a physician about the risks associated with having a public profile and of referencing sex or substance use in their profile. Three months later, networking sites were revisited to see if any changes had been made. The following excerpt is from the paper:

> At baseline, 54.2% of subjects referenced sex and 85.3% referenced substance use on their social networking site profiles. The proportion of profiles in which references decreased to 0 was 13.7% in the intervention group vs. 5.3% in the control group for sex (P =.05) and 26% vs. 22% for substance use (P =.61). The proportion of profiles set to "private" at follow-up was 10.5% in the intervention group and 7.4% in the control group (P = .45). The proportion of profiles in which any of these three protective changes were made was 42.1% in the intervention group and 29.5% in the control group (P =.07).

- **a.** The quote from the paper references four hypothesis tests. For each test, indicate what hypotheses you think were tested and whether or not the null hypothesis was rejected.
- **b.** Based on the information provided by the hypothesis tests, what conclusions can be drawn about the effectiveness of the e-mail intervention?

11.62 The paper **"Children and Adolescents' Exposure to Food and Beverage Marketing in Social Media Apps" (*Pediatric Obesity* [2019]: e12508)** recorded time spent using social media apps for each person in a sample of 38 children age 7 to 11 and a sample of 63 adolescents age 12 to 16. The samples were convenience samples recruited to participate in the study.

The authors of the paper reported the sample means and standard deviations given in the accompanying table. They concluded that, on average, adolescents spent more time than children on social media apps on a typical weekday.

Sample	Sample Size	Sample Mean	Sample Standard Deviation
Children	38	24	31
Adolescents	63	48	55

- **a.** Would it be appropriate to have used a two-sample t test to reach the conclusion that, on average, adolescents spent more time than children on social media apps on a typical weekday?
- **b.** The authors of the paper actually used a different test procedure that does not depend on the assumption of normal population distributions. What additional information would you want to know to determine if it is reasonable to generalize the stated conclusion to all children and adolescents?

Section 11.5 Simulation-Based Inference for Two Means (Optional)

Simulation-based methods for inference about one population mean were introduced in Chapter 10. In this section, simulation-based methods that allow you to test hypotheses and construct confidence intervals for a difference in means are introduced. These methods are especially useful when the conditions of the two-sample methods of Sections 11.1 and 11.2 are not met (when the sample sizes are small and it is not clear that the population distributions are normal).

Simulation-Based Inference About the Difference in Two Treatment Means

Simulation-based randomization tests and bootstrap confidence intervals may be used to learn about the difference in two treatment means using data from an experiment.

Example 11.12 Blue Light Exposure and Blood Glucose Level

● The article **"Bright Light at Night Time Can Seriously Mess with Your Metabolism, Study Finds" (Science Alert, May 20, 2016, sciencealert.com)** describes research conducted to examine the effects of blue light exposure (the type of light emitted by smart phones and computer screens) on a variety of measures, including blood glucose levels and sleepiness. The study was published in the **Public Library of Science ("Morning and Evening Blue-Enriched Light Exposure Alters Metabolic Function in Normal Weight Adults,"** ***PLOS One*** **[2016]: e0155601)**. Adult volunteers of normal weight were randomly assigned to one of two groups. The first group was exposed to blue light for 3 hours in the morning (30 minutes after waking), and the second group was exposed to blue light for 3 hours in the evening (10 hours and 30 minutes after waking).

For each subject, a baseline blood glucose level was obtained 30 minutes before the blue light exposure began. Then blood glucose level was tracked every half hour for 4 hours during and immediately following the blue light exposure, including when the subject ate a meal. One outcome measure of the study was the peak change from baseline in blood glucose level. A negative peak change means that the subject's blood glucose level remained below the baseline measurement during the entire time period. You can perform a hypothesis test using a significance level of 0.05 to determine whether the mean peak change in blood glucose level differs for the A.M. and P.M. blue light exposure groups.

Data for the 9 subjects in the A.M. blue light treatment group and the 10 subjects in the P.M. blue light treatment group are given in the following table.

Subject ID	Group	Glucose Peak Change from Baseline (mg/dL)
1	A.M.	4
2	A.M.	−22
3	A.M.	7
4	A.M.	4
5	A.M.	−8
6	A.M.	−13
7	A.M.	3
8	A.M.	−20
9	A.M.	−10
10	P.M.	28
11	P.M.	12
12	P.M.	12
13	P.M.	33
14	P.M.	25
15	P.M.	22

● **Data set available online**

(continued)

Copyright 2025 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s). Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

Subject ID	Group	Glucose Peak Change from Baseline (mg/dL)
16	P.M.	4
17	P.M.	22
18	P.M.	−2
19	P.M.	−18

The dotplots in Figure 11.2 indicate the assumption that the change distributions are approximately normal is questionable. Because both sample sizes are small, the two-sample *t* test and confidence interval might not be appropriate choices for analyzing the data from this experiment.

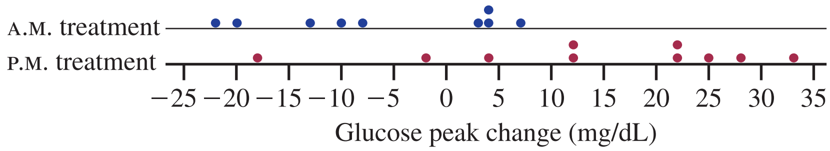


Figure 11.2
Dotplots of peak glucose change for the A.M. and P.M. treatments.

The two treatment group means can be represented as follows:

μ_1 = Mean change in glucose level for the A.M. treatment
μ_2 = Mean change in glucose level for the P.M. treatment

Translating the research question of whether mean change in blood glucose level differs for the A.M. and P.M. blue light exposure treatments results in the following hypotheses:

H_0: $\mu_1 - \mu_2 = 0$
H_a: $\mu_1 - \mu_2 \neq 0$

Summary statistics for the data in the two samples are

	A.M. Treatment	P.M. Treatment
Sample size	9	10
Sample mean	−6.1	13.8
Sample standard deviation	11.02	15.61

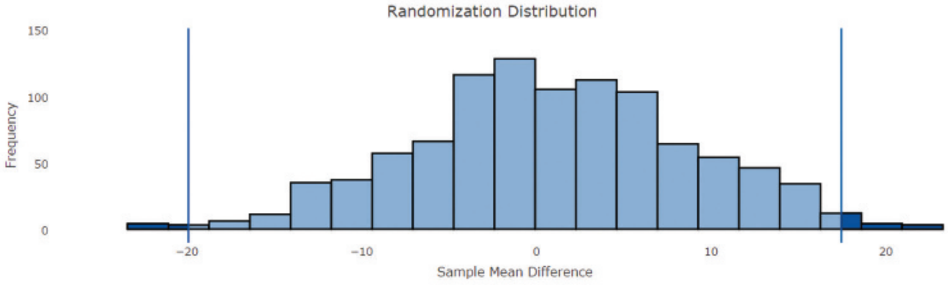
The observed difference in means for change in glucose level, A.M. − P.M., is $\bar{x}_1 - \bar{x}_2 = -19.9$.

A randomization test can be used to determine if there is convincing evidence that the two treatment means differ. In a randomization test using data from an experiment, alternative random assignments of the subjects in a study to two (or more) groups are considered. If the null hypothesis is true, there is no difference in the effect of the two treatments. This means that each subject would have had the same observed change whether they were in the A.M. treatment group or the P.M. treatment group. For example, subject 1, who was in the A.M. treatment group and had a change of 4, would have had a change of 4 even if that subject had been in the P.M. treatment group. If this is the case for every subject, the observed difference in sample means could be due to chance in the random assignment of subjects to treatment groups.

To decide if the observed difference is consistent with what is expected due to chance alone or whether it is evidence of a real difference in treatment means, you begin by exploring what chance differences look like. Simulation is carried out by taking the original 19 observations and randomly assigning them into two groups (one of size 9 and one of size 10) and then calculating a simulated difference in means. This process is repeated many times to form a randomization distribution.

The following figure shows the simulation results from the Shiny app “Randomization Test for Difference in Two Treatment Means” for the difference in means from many different random assignments of the observed data values into two independent groups. This app can be found in the collection that accompanies this text at **www.cengage.com/Peck**.

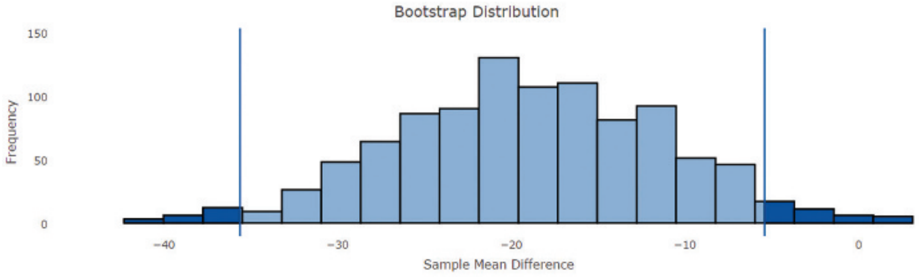
Copyright 2025 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s). Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.



Observed Difference in Sample Means	Estimated Standard Error	Two-Tailed P-value	Number of Values in Tail	Number of Randomizations
-19.911	7.733	0.014	14	1000

The observed difference in means was −19.9. Locating −19.9 in the distribution of simulated differences shows that it would be unusual to observe a difference this extreme if the null hypothesis of no difference in treatment means were true. Based on the randomization distribution, the probability of observing a difference at least as extreme as −19.9 is approximately 0.014. Because this two-sided *P*-value is less than the specified significance level of 0.05, the null hypothesis is rejected. Among many random assignments of the changes in blood glucose levels to hypothetical A.M. and P.M. blue light exposure groups, less than 1.4% produce a difference in means that is at least as inconsistent with the null hypothesis as the observed difference in the means of −19.9. This is evidence that the treatment means are not equal.

The Shiny app "Bootstrap Confidence Interval for Difference in Two Treatment Means" (found in the collection at **www.cengage.com/Peck**) can be used to obtain a 95% confidence interval for the difference in the treatment means by examining the distribution of differences in simulated means calculated from hypothetical randomizations of the combined sample of changes in blood glucose levels into two groups.



Observed Difference in Sample Means	Estimated Standard Error	Lower Bound	Upper Bound	Confidence Level	Number of Differences Outside CI	Number of Randomizations
-19.911	7.699	-35.611	-5.411	95%	48	1000

The given Shiny app output provides the distribution for 1000 simulated differences in the means for alternative random assignments of the changes in blood glucose levels to two groups. Identifying the 2.5% of differences in means on both the lowest and highest ends of the distributions provides the endpoints for a 95% bootstrap confidence interval.

For this simulation, the 95% bootstrap confidence interval for the difference in treatment means is (−35.6, −5.4). You can be 95% confident that the actual difference in the mean change in blood glucose levels, A.M. – P.M., falls between −35.6 and −5.4. Notice that this confidence interval does not include zero, and this is consistent with rejecting the null hypothesis that the difference in the mean change in glucose level for the two treatments is zero. With both endpoints negative, you would estimate that the mean change is greater for the P.M. blue light treatment than for the A.M. blue light treatment by somewhere between 5.4 and 35.6 mg/dL.

Copyright 2025 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s). Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

Simulation-Based Inference for the Difference in Two Population Means Using Independent Samples

Sometimes you want to compare the means for two populations using independent samples selected at random from the populations, in contrast to using data from an experiment where subjects are assigned at random to treatment groups. The simulation-based methods for comparing two means using independent random samples are a little different from the methods used with data from an experiment.

Example 11.13 Freshman Year Weight Gain Revisited

● Example 11.4 described a study of freshman year weight gain **("Predicting the 'Freshman 15': Environmental and Psychological Predictors of Weight Gain in First-Year University Students," *Health Education Journal* [2016]: 321–332)**. In that study, weight gain (in kilograms) during the freshman year was recorded for independent random samples of first-year students who lived on campus and first-year students who lived off campus. The data from Example 11.4 are also given here.

On Campus	Off Campus
2.0	1.6
2.3	3.1
1.1	−2.8
−2.0	0.0
−1.9	0.2
5.6	2.9
2.6	−0.9
1.1	3.8
5.6	0.7
8.2	−0.1

These samples are quite small, but the boxplots of the data from the two samples were reasonably symmetric and there were no outliers, so a two-sample *t* confidence interval was used to estimate the difference in population means. However, you might still choose to use a simulation-based method that does not depend on the assumption of normal population distributions.

The means for the populations of first-year students living on and off campus were represented by

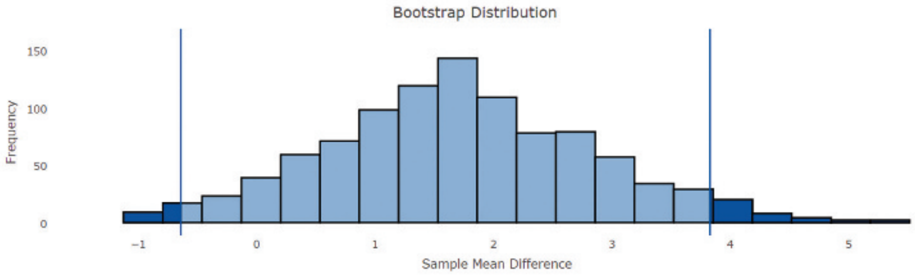
μ_1 = population mean weight gain for first-year students living on campus
μ_2 = population mean weight gain for first-year students living off campus

In the case of sampling from two populations, you begin by assuming that each sample is representative of the population from which it was selected. Then, as was the case in the one-sample situation, a bootstrap sample from each population is selected using random sampling, with replacement, from each of the two original samples. The difference in the means of these two bootstrap samples is calculated. This process is then repeated many times to form a bootstrap distribution for the difference in sample means.

After a bootstrap distribution of simulated differences in sample means is constructed, it can be used to produce a 95% confidence interval for the difference in the population means. The endpoints of the confidence interval are the value with 2.5% of the simulated differences in the bootstrap distribution below and the value with 2.5% of the simulated differences in the bootstrap distribution above. Locating these differences in the bootstrap distribution on the next page gives a 95% confidence interval for the difference in mean weight gain for students living on campus and students living off campus of (−0.64, 3.83). You can be 95% confident that the difference in population mean weight gains, On Campus – Off Campus, falls between −0.64 and 3.83 kg. These values are not very different from the endpoints of the two-sample *t* confidence interval from Example 11.4, which was (−0.98, 4.20).

● **Data set available online**

Here is the output from the Shiny app “Bootstrap Confidence Interval for the Difference in Two Population Means Using Independent Samples.” This app can be found in the collection at **www.cengage.com/Peck**.



Observed Difference in Sample Means	Estimated Standard Error	Lower Bound	Upper Bound	Confidence Level	Number of Differences Outside CI	Number of Randomizations
1.610	1.129	-0.640	3.830	95%	50	1000

It is also possible to compare the population mean weight gains for students living on and off campus using a simulation-based hypothesis test, but the simulation is carried out differently than for the confidence interval because the null hypothesis states that the population mean weights are equal without specifying their values. Suppose that you want to test the following hypotheses:

$$H_0: \mu_1 - \mu_2 = 0$$
$$H_a: \mu_1 - \mu_2 \neq 0$$

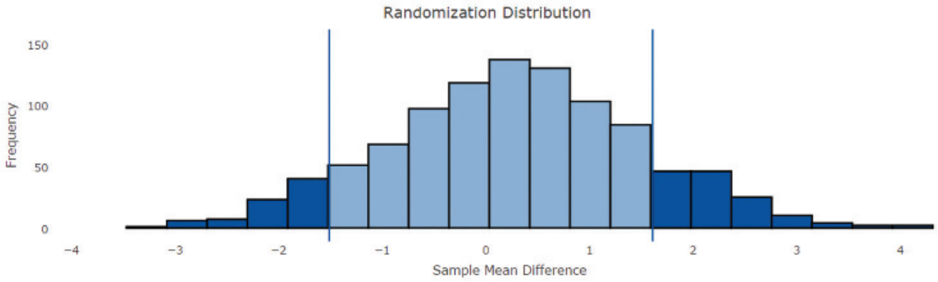
Summary statistics for the weight gains (in kg) calculated using the given data for students living on campus and for those living off campus follow.

	On Campus	Off Campus
Sample size	10	10
Sample mean	2.46	0.85
Sample standard deviation	3.26	2.03

The observed difference in the means for On Campus − Off Campus is $\bar{x}_1 - \bar{x}_2 = 2.46 - 0.85 = 1.61$ kg.

To create a randomization distribution, the two samples are combined into one, and then random samples are selected with replacement from the combined sample. This creates a randomization distribution of the difference in sample means that is consistent with the null hypothesis of no difference in population means.

One example of a randomization distribution based on 1000 simulated differences in means using the Shiny app “Randomization Test for the Difference in Two Population Means Using Independent Samples” is shown here. This app can be found in the collection at **www.cengage.com/Peck**.



Observed Difference in Sample Means	Estimated Standard Error	Two-Tailed P-value	Number of Values in Tail	Number of Simulated Differences
1.610	1.218	0.220	220	1000

Copyright 2025 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s). Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

The randomization distribution in this case represents 1000 differences in two independent sample means, each calculated from two random samples of size 10 taken with replacement from the combined sample of 20 weight gains from students living on and off campus.

Using the randomization distribution from this simulation, the probability of observing a difference in the sample means at least as extreme as 1.61 is 0.220. This two-tailed *P*-value is larger than the 0.05 significance level. You would fail to reject the null hypothesis, and conclude that there is not convincing evidence of a difference in the population means for weight gains of students living on campus and students living off campus. Notice that this is consistent with zero being included in the observed bootstrap confidence interval.

Simulation-Based Inference About Two Population Means Using Paired Samples

In Section 11.2, when the samples are paired, hypothesis tests and confidence intervals for the difference in two population means involved calculating the sample differences and then using one-sample methods with the sample of differences. This is also the case for simulation-based methods—you calculate the sample differences and then use the one-sample bootstrap confidence interval and simulation-based randomization test that were introduced in Chapter 10.

Example 11.14 Two Ways to Measure Body Fat Mass

● The authors of the paper **"Body Composition Measurement in Young Children Using Quantitative Magnetic Resonance: A Comparison with Air Displacement Plethysmography" (*Pediatric Obesity* [2018]: 365–373)** compared two different methods for measuring body fat mass. Body fat mass was measured for a representative sample of 5-year-old children using two different methods. One method was a new method using magnetic resonance (MR) and the other was a commonly used air displacement method (AD). The new MR method is faster and does not require subjects to remain motionless, making it an attractive alternative for use with children as long as the measurements are reliable. Because the samples were paired, the differences in body fat mass measurements (in kg, AD minus MR) were calculated. The differences for 12 children (a subset of the data given in a graph that appeared in the paper) are given in the accompanying table.

Child ID	1	2	3	4	5	6	7	8	9	10	11	12
Difference in Body Fat Measurement (AD minus MR)	1.3	0.4	−0.6	1.5	−0.3	0.3	−0.3	0.4	−1.1	1.6	0.5	−1.2

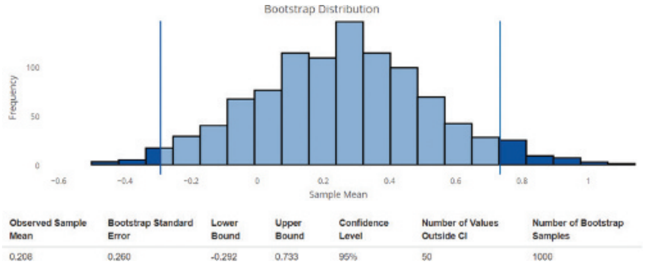
The sample size here is small, so you might be hesitant to use the one-sample *t* test for the paired sample of differences. Instead, you may choose to use the simulation-based methods presented in Section 10.8. To construct a confidence interval for the mean difference in body mass measurement, begin by selecting bootstrap samples, with replacement, from the original sample of differences. Here is one bootstrap sample:

Child ID	12	8	10	4	6	2	5	8	9	11	2	12
Difference in Body Fat Measurement (AD minus MR)	−1.2	0.4	1.6	1.5	0.3	0.4	−0.3	0.4	−1.1	0.5	0.4	−1.2

The sample mean for this bootstrap sample is $\bar{x}_d = 0.142$. You can repeat this process many times, and the resulting bootstrap distribution of $\bar{x}_d$ values provides information about sampling variability that can be used to find a confidence interval for the population mean difference, μ_d, where μ_d is the mean difference calculated as

(mean body mass for AD method) − (mean body mass for MR method)

The Shiny app "Bootstrap Confidence Interval for One Mean" (which can be found in the App collection at **www.cengage.com/Peck**) produces a bootstrap distribution that can be used to obtain a bootstrap confidence interval for the mean difference, as shown here.



Observed Sample Mean	Bootstrap Standard Error	Lower Bound	Upper Bound	Confidence Level	Number of Values Outside CI	Number of Bootstrap Samples
0.208	0.260	-0.292	0.733	95%	50	1000

For this particular simulation, the bootstrap method produces a 95% confidence interval for the population mean difference, μ_d, of −0.292 to 0.733. You can be 95% confident that this interval contains the population mean difference.

You may also wish to perform a hypothesis test to determine if these data provide convincing evidence that the mean body fat mass measurement differs for the two methods. This would be a test of the claim that the mean difference is not equal to 0. To do this, the data values in the original ample can be shifted to represent a sample from a hypothetical population with mean $\mu_d = 0$ by subtracting 0.208 (the mean of the original sample of differences) from each difference, as follows:

Child ID	1	2	3	4	5	6	7	8	9	10	11	12
Difference	1.3	0.4	−0.6	1.5	−0.3	0.3	−0.3	0.4	−1.1	1.6	0.5	−1.2
Difference − 0.208	1.092	0.192	−0.808	1.292	−0.508	0.092	−0.508	0.192	−1.308	1.392	0.292	−1.408

The relevant hypotheses are:

$$H_0\colon \mu_d = 0$$
$$H_a\colon \mu_d \neq 0$$

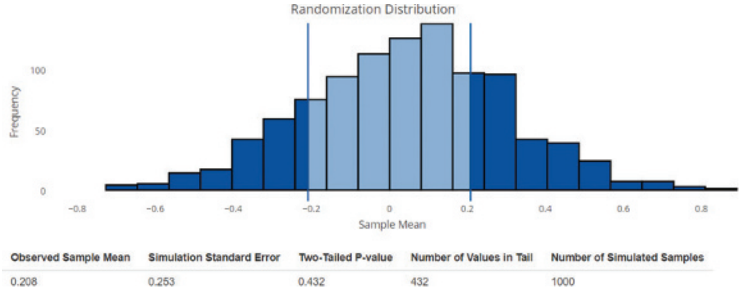
The simulation approach begins with resampling from the sample created by shifting the original sample of differences to have a mean of 0. Here is one simulated sample from the shifted sample:

Child ID	6	9	2	1	11	3	4	6	9	5	1	11
Difference − 0.208	0.092	−1.308	0.192	1.092	0.292	−0.808	1.292	0.092	−1.308	−0.508	1.092	0.292

The sample mean for this sample is $\bar{x}_d = 0.042$.

A randomization distribution is generated by selecting many simulated samples from the shifted sample. This distribution can be viewed as the distribution of the sample mean difference if $\mu_d = 0$ is true. A *P*-value is the probability of obtaining a value at least as inconsistent with H_0 as what actually resulted. This means that the *P*-value for testing the null hypothesis $H_0\colon \mu_d = 0$ against the two-sided alternative $H_a\colon \mu_d \neq 0$ is two times the probability of observing $\bar{x}_d \geq 0.208$ when H_0 is true. This probability can be approximated using the proportion of simulated values of $\bar{x}_d$ that fall at or above 0.208 in the distribution generated using the simulated samples.

The following randomization distribution was produced using the Shiny app "Randomization Test for One Mean." This app can be found at **www.cengage.com/Peck**.

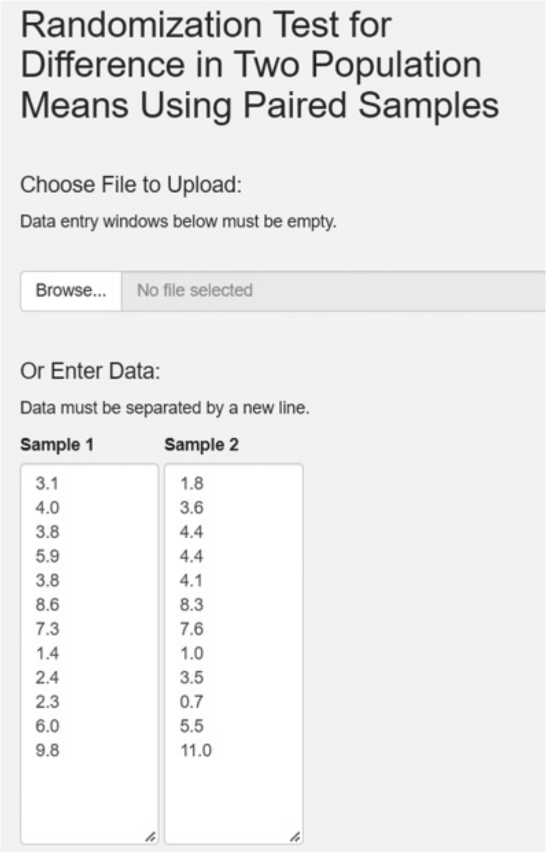


Observed Sample Mean	Simulation Standard Error	Two-Tailed P-value	Number of Values in Tail	Number of Simulated Samples
0.208	0.253	0.432	432	1000

Copyright 2025 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s). Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

The two-tailed *P*-value based on this simulation is 0.432. This *P*-value is greater than the 0.05 significance level, and so you fail to reject H_0. There is not convincing evidence that the mean difference is different from zero. Notice that that this conclusion is consistent with the confidence interval calculated above, which included 0 in the interval.

To use the one-sample Shiny apps with paired samples, you need to first calculate the sample differences. There are also two apps available that allow you to enter the data from the two samples without first having to calculate the differences. These apps are "Bootstrap Confidence Interval for the Difference in Two Population Means Using Paired Samples" and "Randomization Test for Difference in Two Population Means Using Paired Samples." For example, the following output was produced by the Shiny app "Randomization Test for Difference in Two Population Means Using Paired Samples" using the data from Example 11.14.



Randomization Test for Difference in Two Population Means Using Paired Samples

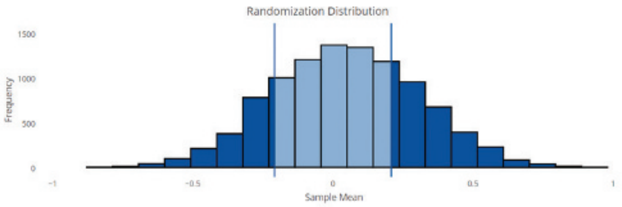
Choose File to Upload:

Data entry windows below must be empty.

Or Enter Data:

Data must be separated by a new line.

Sample 1	Sample 2
3.1	1.8
4.0	3.6
3.8	4.4
5.9	4.4
3.8	4.1
8.6	8.3
7.3	7.6
1.4	1.0
2.4	3.5
2.3	0.7
6.0	5.5
9.8	11.0



Observed Sample Mean Difference	Estimated Standard Error	Two-Tailed P-value	Number of Values in Tail	Number of Simulated Samples
0.208	0.260	0.438	4382	10000

Notice that the *P*-value based on this randomization distribution is 0.438, compared to the *P*-value of 0.432 in Example 11.14. Because simulation-based methods, such as randomization tests, are based on random sampling, the randomization distribution will vary from simulation to simulation, depending on the outcomes of the sampling process. But usually the estimated *P*-values (or the bootstrap confidence interval endpoints) are similar from one simulation to another (as was the case in this example) as long as the number of samples used in the simulation is large.

Exercises 11.63–11.70

● **Data set available online**

11.63 ● The Sheboygan (Wisconsin) Fire Department received a report on the potential effects of reductions in the number of firefighters it employs (**"Study of Fire Department Causes Controversy,"** ***USA TODAY NETWORK-Wisconsin*****, December 22, 2016**).

In one section of the report, the average working heart rate percentage (the percentage value of the observed maximum heart rate during firefighting drills divided by an age-adjusted maximum heart rate for each firefighter) was reported for the driver of the first-arriving fire engine when only two firefighters (including the driver) were present, and for the driver of the first-arriving fire engine when more than two firefighters (up to five) were present.

The average working heart rate percentages were based on an earlier study, which included data from a sample of six drills using only two firefighters and from a sample of 18 drills using more than two firefighters. For purposes of this exercise, assume that these samples are representative of all drills with two firefighters and all drills with more than two firefighters.

The following data values are consistent with summary statistics given in the paper. Do these data support the claim that the mean average working heart rate percentage for the driver in a two-firefighter team is greater than the mean average working heart rate percentage for the driver in teams containing from three to five firefighters? Use a 0.05 significance level to carry out a randomization test of the given claim. You can make use of the Shiny apps in the collection at **www.cengage.com/Peck**.

Working Heart Rate Percentage	
Two Firefighters	**More than Two Firefighters**
97.7	73.6
87.7	63.7
89.5	72.0
97.3	83.1
81.9	54.2
79.7	73.1
	75.2
	84.5
	72.8
	70.0
	87.1
	81.1
	57.5
	68.4
	80.6
	57.8
	79.5
	75.0

11.64 Use the information given in the previous exercise to construct a 95% bootstrap confidence interval to estimate the difference in mean average working heart rates for the driver in teams of two firefighters and the driver in teams of from three to five firefighters. Interpret the interval in context. You can make use of the Shiny apps in the collection at **www.cengage.com/Peck**.

11.65 ● Studies have been conducted to evaluate the effectiveness of psilocybin mushrooms on improving the quality of life for patients with cancer (**"A Dose of a Hallucinogen from a 'Magic Mushroom,' and Then Lasting Peace,"** ***The New York Times*****, December 1, 2016**). In one study, patients were randomly assigned to either a low-dose psilocybin treatment or to a high-dose treatment. One outcome that was measured was a "Personal Meaning" score, collected 5 weeks after the psilocybin treatment. Higher scores indicated greater "Personal Meaning." The following data were estimated from a graphical display in the article. Do these data support the claim that the mean Personal Meaning score for patients with cancer taking a high dose of psilocybin is greater than the mean Personal Meaning score for patients with cancer taking a low dose? Use a randomization test to answer this question. You can make use of the Shiny apps in the collection at **www.cengage.com/Peck**.

Personal Meaning Low Dose		Personal Meaning High Dose	
3	1	1	7
2	2	5	7
6	4	7	8
4	4	7	8
4	2	5	8
2		5	7
3		5	6
7		6	4
7		6	7
6		7	7
6		6	7
7		7	6
5		5	

11.66 Use the information in the previous exercise to construct a 95% bootstrap confidence interval to estimate the difference in mean Personal Meaning scores for patients with cancer in the high-dose and low-dose psilocybin groups. Interpret the interval in context. You can make use of the Shiny apps in the collection at **www.cengage.com/Peck**.

11.67 ● A new set of cognitive training modules called ONTRAC was developed to help children with attention deficit hyperactivity disorder (ADHD) to improve focus and to more easily dismiss distractions **("Training Sensory Signal-to-Noise Resolution in Children with ADHD in a Global Mental Health Setting," *Translational Psychiatry*, April 12, 2016, nature.com)**. Eighteen children with ADHD were randomly assigned to one of two treatment groups. One group of 11 children received the ONTRAC treatment and another group of 7 children received a control treatment.

Values for 1-year improvement in ADHD Severity Score consistent with graphs and summary statistics in the research article appear in the following table.

ONTRAC	Control
11.0	4.3
8.8	3.8
12.1	8.5
10.9	4.7
8.0	1.5
8.2	7.7
11.9	8.0
11.1	
11.8	
10.8	
12.0	

a. Explain why it is wise to be wary of using the two-sample t methods to analyze the data from this study.

b. Do these data support the claim that the mean 1-year improvement in ADHD Severity Score for the ONTRAC treatment is different from the mean 1-year improvement in ADHD Severity Score for the control treatment? Use a randomization test with significance level 0.05 to answer this question. You can make use of the Shiny apps in the collection at **www.cengage.com/Peck**.

11.68 Use the information in the previous exercise to construct a 95% bootstrap confidence interval to estimate the difference in mean 1-year improvement in ADHD Severity Score for the ONTRAC treatment and the control treatment. Interpret the interval in context. You can make use of the Shiny apps in the collection at **www.cengage.com/Peck**.

11.69 ● New "closed loop" (CL) devices have been developed to help to suppress overactive brain activity in patients with conditions such as Parkinson's disease and epilepsy **("Conceptualization and Validation of an Open-Source Closed-Loop Deep Brain Stimulation System in Rats," *Scientific Reports*, April 21, 2015, nature.com)**. The CL device is implanted directly into a specific area of the brain, and one of the key advantages is that it can immediately apply a treatment in reaction to heightened brain activity. This may help to reduce the duration of seizures and other periods of uncontrolled movement in patients.

A study was conducted on the effectiveness of the CL device on brain activity in rats. First, each of seven rats was observed for 15 minutes with the CL device implanted but not activated (OFF). The percentage of the time that each rat was moving was recorded. Then, after each CL device was activated (CL), the percentage of the time that the rat was moving was recorded over another 15-minute period.

The OFF and CL data values are approximated from a graph in the research article and appear in the table below, along with the calculated difference in movement, OFF − CL.

Rat ID	OFF	CL	Difference (OFF − CL)
1	98.8	56.6	42.2
2	56.9	25.5	31.4
3	56.7	40.0	16.7
4	58.1	68.1	−10.0
5	50.8	45.1	5.7
6	53.1	35.7	17.4
7	56.8	44.5	12.3

a. Explain why two-sample t methods may not be appropriate in this context.

b. For purposes of this exercise, assume that these rats are representative of rats in general. Do these data support the claim that the mean difference in movement, OFF − CL, is greater than zero? Carry out a randomization test to answer this question. You can make use of the Shiny apps in the collection at **www.cengage.com/Peck**.

11.70 Use the information given in the previous exercise to calculate a 95% bootstrap confidence interval to estimate the mean difference in movement, OFF − CL. Interpret the interval in context. You can make use of the Shiny apps in the collection at **www.cengage.com/Peck**.

Section 11.6 Simulation-Based Inference for Two Proportions (Optional)

The large-sample methods for estimating the difference between two proportions and for testing hypotheses about the difference between two proportions (Section 11.3) require independent random samples and sample sizes that are large enough to ensure that the sampling distribution of the differences in the sample proportions, $\hat{p}_1 - \hat{p}_2$, is approximately normal. When one or both sample sizes are not large enough, simulation-based methods can be used to estimate the difference in two proportions or to test hypotheses about the difference in two proportions.

Bootstrap Confidence Intervals for the Difference Between Two Population Proportions or Two Treatment Proportions

Example 11.15 Anti-Clotting Medications After Hip or Knee Surgery

In the study described in the article **"A Pilot Study Comparing Hospital Readmission Rates in Patients Receiving Rivaroxaban or Enoxaparin After Orthopedic Surgery," (ptcommunity.com, 2016)**, researchers identified patients who had received hip or knee replacement surgery and who had been given one of two types of medication to prevent blood clotting afterward. This was an observational study, conducted by reviewing the medical charts for patients who had received hip or knee replacement surgery. The outcome of interest was whether the patient was readmitted to the hospital within 30 days after leaving the hospital after surgery. The researchers were interested in estimating the difference in the proportions of patients readmitted for the two drugs (rivaroxaban and enoxaparin).

The study found that 8 of 213 patients who had been given rivaroxaban to prevent blood clots were readmitted to the hospital within 30 days after their surgeries, and that 1 out of 27 patients who had taken enoxaparin to prevent clotting was readmitted within 30 days. Based on the sample data, what can you learn about the difference in the proportion of patients who receive rivaroxaban who are readmitted and the corresponding proportion for patients who receive enoxaparin?

You want to estimate the difference in the proportion of patients readmitted to the hospital for those who receive rivaroxaban and the proportion readmitted for those who receive enoxaparin. If p_1 is the proportion of readmissions for patients who take rivaroxaban and p_2 is the proportion of readmissions for patients who take enoxaparin, then you can estimate the difference, $p_1 - p_2$.

Only 8 patients who received rivaroxaban were readmitted, and only 1 of the patients who received enoxaparin was readmitted. The sample sizes are too small to use the large-sample z confidence interval for a difference in population proportions. Instead, you can estimate the difference in population proportions, $\hat{p}_1 - \hat{p}_2$, using a bootstrap simulation-based method. For this example, a 95% confidence level will be used.

The sample of patient records was not randomly selected, but the researchers believed it to be representative of the population of patients undergoing hip and knee replacements who receive rivaroxaban and the population of patients who received enoxaparin. Based on the given information, you know

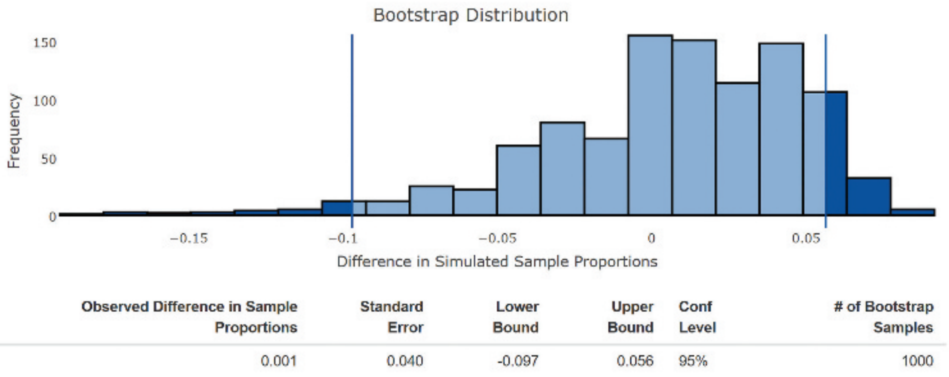
$$\begin{array}{ll} n_1 = 213 & n_2 = 27 \\ \hat{p}_1 = 8/213 = 0.038 & \hat{p}_2 = 1/27 = 0.037 \end{array}$$

To generate bootstrap simulated differences, $\hat{p}_1 - \hat{p}_2$, first consider a hypothetical population in which the proportion of success is 0.038 and take a sample of size 213 from this hypothetical population. This simulated sample produces a simulated value of $\hat{p}_1$. Then consider a hypothetical population in which the proportion of success is 0.037 and take a sample of size 27. This simulated sample produces a simulated value of $\hat{p}_2$. This pair of values, $\hat{p}_1$ and $\hat{p}_2$, produces a bootstrap simulated value of $\hat{p}_1 - \hat{p}_2$.

Copyright 2025 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s). Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

For example, a simulated value of $\hat{p}_1 = 11/213 = 0.052$ was generated from a hypothetical population with population proportion of successes $p_1 = 0.038$. A simulated value of $\hat{p}_2 = 1/27 = 0.037$ was generated from a hypothetical population with population proportion of successes $p_2 = 0.037$. The simulated difference in sample proportions is $\hat{p}_1 - \hat{p}_2 = 0.052 - 0.037 = 0.015$.

A graph produced by the Shiny app "Bootstrap Confidence Interval for Difference in Two Proportions" of 1000 simulated differences in sample proportions is shown below. (This Shiny app can be found in the collection at **www.cengage.com/Peck**.)



Observed Difference in Sample Proportions	Standard Error	Lower Bound	Upper Bound	Conf Level	# of Bootstrap Samples
0.001	0.040	-0.097	0.056	95%	1000

The smallest 2.5% of the simulated differences in sample proportions were −0.097 or less, and the largest 2.5% of the simulated differences were 0.056 or greater. The 95% confidence interval based on the bootstrap resampling is (−0.097, 0.056).

If the samples are representative of the two populations of interest (patients that take rivaroxaban and patients who take enoxaparin), you can be 95% confident that the actual difference in the proportion of patients readmitted to the hospital after taking rivaroxaban and the proportion readmitted after taking enoxaparin is between −0.097 and 0.056. Zero is contained in this interval, indicating that it is plausible that there is no real difference in the readmission proportions for the two medications.

Randomization Tests for the Difference Between Two Population Proportions or Two Treatment Proportions

In some cases, you want to determine if two population proportions or two treatment proportions differ, or if one is smaller or larger than the other. In these situations, you would test hypotheses about the difference in the proportions.

Example 11.16 Oxytocin Nasal Spray and Social Interaction

Oxytocin is a synthetic hormone that may improve social interaction for young children with autism. An experiment was conducted to evaluate if the use of oxytocin delivered through a nasal spray improves social interaction experiences (**"The Effect of Oxytocin Nasal Spray on Social Interaction Deficits Observed in Young Children with Autism,"** ***Molecular Psychology*** **[2016]: 1225–1231**).

In many studies, participants drop out for one reason or another. Sometimes participants move to a new location, sometimes they change their minds about participating, and they always have the right to just discontinue participation in the study.

One concern in the oxytocin nasal spray and autism study was that young children with different diagnoses might drop out at different rates. In fact, 4 out of 23 young children with autism spectrum disorder dropped out of the study, and 4 out of 16 young children with pervasive developmental disorder dropped out. The researchers assumed that the young children in the study were representative of the population of all young children who might experience improved social interactions after taking oxytocin nasal spray.

Copyright 2025 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s). Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

Do these data provide evidence that the proportion who drop out is different for the two different types of autism?

Here is a summary of the information in this example:

Diagnosis	Population Proportion	Sample Size	Sample Proportion
Autism spectrum disorder	p_1 = proportion of all young children with autism spectrum disorder who would drop out of the oxytocin study	$n_1 = 23$	$\hat{p}_1 = 0.174$
Pervasive developmental disorder	p_2 = proportion of all young children with pervasive developmental disorder who would drop out of the oxytocin study	$n_2 = 16$	$\hat{p}_2 = 0.250$

To answer the question of interest, you can test the null hypothesis that the two population proportions are equal. The population proportions to be compared are p_1 = proportion of all young children with autism spectrum disorder who would drop out of the oxytocin study and p_2 = proportion of all young children with pervasive developmental disorder who would drop out of the oxytocin study. The null hypothesis is $H_0: p_1 - p_2 = 0$. Initially, the researchers did not know which diagnosis group might have a larger dropout rate, and so the alternative hypothesis is two-sided, $H_a: p_1 - p_2 \neq 0$.

A large-sample *z* test for a difference in population proportions is not appropriate because the sample sizes are small. The number of young children who dropped out of the study was 4 in each of the two diagnosis groups. Because the sample size conditions for the two-sample *z* test are not met, a simulation-based randomization test can be used instead of a large-sample test. For this example, a significance level of $\alpha = 0.05$ will be used.

The observed difference in the sample proportions is $\hat{p}_1 - \hat{p}_2 = 0.174 - 0.250 = -0.076$. Recall that in the simulation-based test for a difference in two population means (Section 11.5), data from the two samples were combined into one group to represent the possibility that the two samples might have been taken from the same population. A similar process is used for the simulation-based test for the difference in two population proportions. The two samples are combined into one, and then random samples are selected with replacement from the combined samples (which represents the common population) to create a simulation-based distribution of the differences in sample proportions that is consistent with the null hypothesis of no difference in population or treatment proportions.

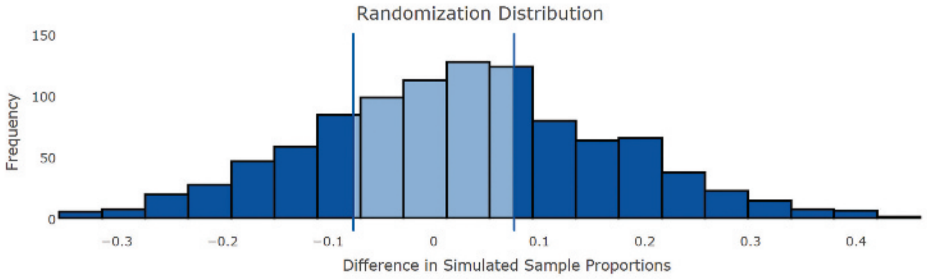
This resampling approach to a test for two proportions involves pooling the two original samples together into one hypothetical collection of $23 + 16 = 39$ young children in which $4 + 4 = 8$ are identified as dropping out of the study. The proportion of children who would drop out based on this hypothetical collection is $\hat{p} = 8/39 = 0.205$. This is known as the pooled sample proportion.

A sample of $n_1 = 23$ is drawn from the combined, or "pooled," collection, with replacement, and a simulated value of $\hat{p}_1$ is calculated. A sample of $n_2 = 16$ is also drawn from the pooled collection, again with replacement, and a simulated value of $\hat{p}_2$ is calculated. These are the values needed to calculate a simulated difference $\hat{p}_1 - \hat{p}_2$. This process is repeated many times in order to construct the randomization distribution.

For example, suppose a sample of $n_1 = 23$ young children drawn from a hypothetical population with pooled proportion $\hat{p} = 0.205$ results in $\hat{p}_1 = 6/23 = 0.261$, and a separate sample of $n_2 = 16$ children drawn from the same population with pooled proportion $\hat{p} = 0.205$ results in $\hat{p}_2 = 4/16 = 0.250$. The simulated difference in the sample proportions is $\hat{p}_1 - \hat{p}_2 = 0.261 - 0.250 = 0.011$.

A randomization distribution for the difference in two proportions was produced using the Shiny app "Randomization Test for Two Proportions" and is shown at the top of the next page. (This Shiny app can be found in the collection at **www.cengage.com/Peck**.)

Copyright 2025 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s). Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.



Observed Difference in Sample Proportions	Standard Error	P-Value	# of Simulated Samples	# of Values in Tail
-0.076	0.138	0.594	1000	594

Because the alternative hypothesis in this example is two-sided, the *P*-value is the sum of the proportion of simulated differences in the two regions representing simulated differences that are as extreme or more extreme than the observed difference in sample proportions. In this example, the observed difference in sample proportions was -0.076, so the *P*-value is the sum of the proportion of simulated differences that are less than -0.076 and the proportion of simulated differences that are greater than 0.076. For the randomization distribution for this particular simulation, this is 0.594. Because this *P*-value is greater than the specified significance level $\alpha = 0.05$, you would fail to reject the null hypothesis.

Based on the sample data, there is not convincing evidence that there is a difference in dropout proportions for young children with a diagnosis of autism spectrum disorder and for children with a diagnosis of pervasive developmental disorder.

Exercises 11.71–11.78

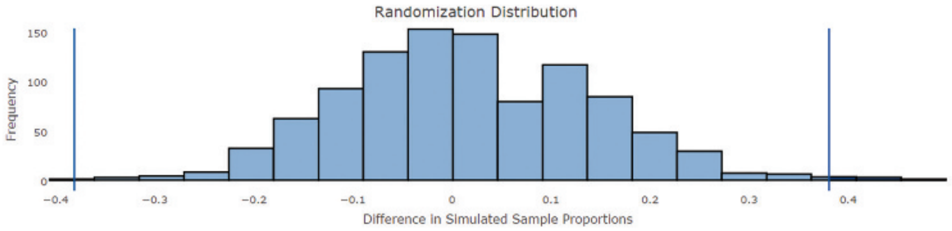
11.71 The article **"Rapid Evolutionary Response to a Transmissible Cancer in Tasmanian Devils" (nature.com)** describes the spread of devil facial tumor disease (DFTD), which is a fatal form of cancer that swept through the Tasmanian devil population near the beginning of the 21st century. Researchers studied the genetic reaction of the Tasmanian devils by comparing the rates of occurrence of specific genetic markers of interest before and after DFTD swept across the island.

One region of Tasmania is called West Pencil Pine. Analysis of 21 tissue specimens taken from a representative sample of Tasmanian devils living in West Pencil Pine in 2006, before DFTD swept through, revealed that 5% had a specific genetic marker. Also analyzed were 42 tissue specimens from a representative sample of devils living in the same region in 2013 and 2014, after DFTD. In this sample, 43% had the same genetic marker. A significant and substantial change in these rates would indicate a remarkably fast evolution in the genetic code of the Tasmanian devils to protect against DFTD.

a. Explain why the data from this study should not be analyzed using a large-sample hypothesis test for the difference in two population proportions.

b. Use the output at the bottom of the page from the Shiny app "Randomization Test for Two Proportions" to carry out a hypothesis test to determine if there is convincing evidence that the proportion

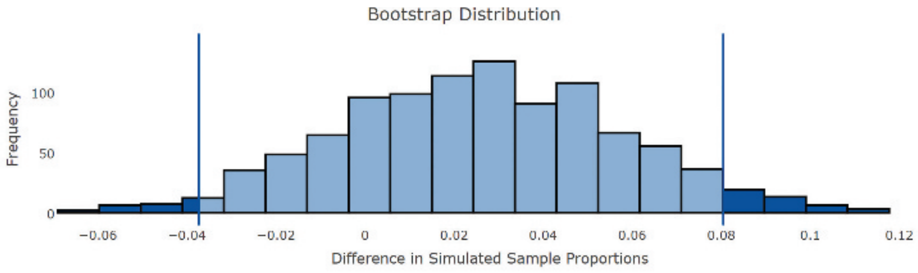
Output for Exercise 11.71(b)



Observed Difference in Sample Proportions	Standard Error	P-Value	# of Simulated Samples	# of Values in Tail
-0.381	0.121	0.002	1000	2

Copyright 2025 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s). Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

Output for Exercise 11.71(c)



Observed Difference in Sample Proportions	Standard Error	Lower Bound	Upper Bound	Conf Level	# of Bootstrap Samples
0.022	0.031	-0.037	0.080	95%	1000

of Tasmanian devils with the genetic marker was greater after DFTD than before DFTD.

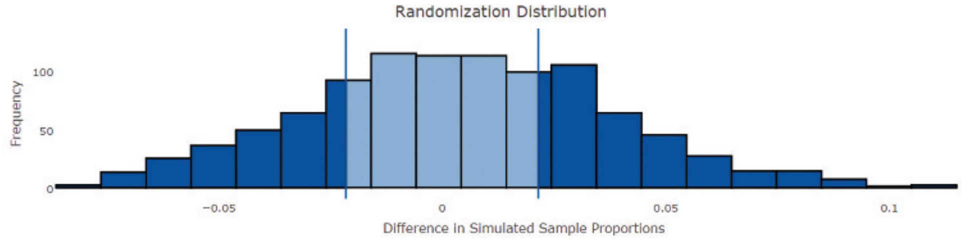
c. Use the output above from the Shiny app "Bootstrap Confidence Interval for Difference in Two Proportions" to identify a 95% confidence interval for the difference in the rates of occurrence of the specific genetic marker in the genes of Tasmanian devils, before and after DFTD. Interpret the confidence interval in context.

11.72 The report **"The 2016 Inside Higher Ed Survey of Faculty Attitudes on Technology" (insidehighered.com)** describes the results of a survey of 1129 full-time college faculty and 293 part-time college faculty. Survey participants were asked if they require undergraduate students to submit papers through plagiarism-detection software; 40% of the full-time faculty and 38% of the part-time faculty said "yes." Notice that the sample sizes for the two groups—full-time college faculty and part-time college faculty—are large enough to satisfy the conditions for a large-sample test and a large-sample confidence interval for a difference in two population proportions. Even though the sample sizes are large, simulation-based methods can still be used.

a. Use the output at the bottom of the page from the Shiny app "Randomization Test for Two Proportions" to carry out a randomization test to determine if the proportion who require students to submit papers through plagiarism-detection software is different for full-time faculty and part-time faculty.

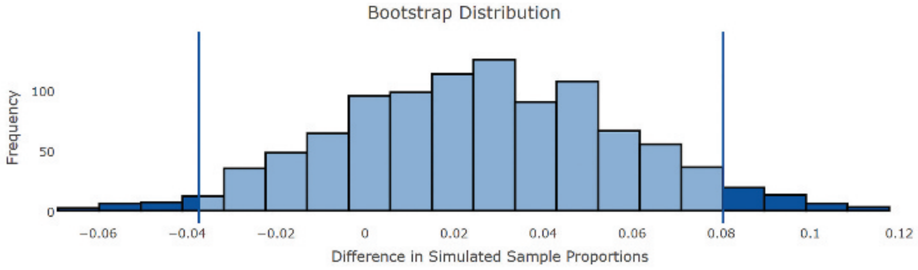
b. Use the output at the bottom of the page from the Shiny app "Bootstrap Confidence Interval for Difference in Two Proportions" to identify a 95% confidence interval for the difference in the population proportions of full-time college faculty and part-time college faculty who require students to submit papers through plagiarism-detection software. Interpret the confidence interval in context.

Output for Exercise 11.72 (a)



Observed Difference in Sample Proportions	Standard Error	P-Value	# of Simulated Samples	# of Values in Tail
0.022	0.034	0.502	1000	502

Output for Exercise 11.72 (b)



Observed Difference in Sample Proportions	Standard Error	Lower Bound	Upper Bound	Conf Level	# of Bootstrap Samples
0.022	0.031	-0.037	0.080	95%	1000

Copyright 2025 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s). Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

11.73 Researchers were interested in comparing regular-intensity exercise and high-intensity exercise for patients recovering from hospitalization due to chronic obstructive pulmonary disease (COPD). The researchers followed patients in Denmark who were enrolled in each of the two types of exercise programs (**"Increased Mortality in Patients with Severe COPD Associated with High-Intensity Exercise: A Preliminary Cohort Study,"** ***Journal of Chronic Obstructive Pulmonary Disease*** **[2016]: 2329–2334)**. Each exercise program lasted for 8 weeks. The patients were followed for a total of 1.5 years. The researchers observed that 5 out of the 15 patients in the high-intensity group died within a year and a half, but that none of the 16 patients in the regular-intensity group died within a year and a half.

a. Explain why the data from this study should not be analyzed using a large-sample hypothesis test for a difference in two population proportions.

b. Carry out a hypothesis test to determine if there is convincing evidence of a difference in the population proportions who die within 1.5 years for the two exercise programs. Use the Shiny app "Randomization Test for Two Proportions" (in the collection at **www.cengage.com/Peck**) to report an approximate P-value and then use it to reach a decision in the hypothesis test. Remember to interpret the results of the test in context.

c. Use the Shiny app "Bootstrap Confidence Interval for Difference in Two Proportions" (in the collection at **www.cengage.com/Peck**) to obtain a 95% bootstrap confidence interval for the difference in the population proportions in patients who die within 1.5 years for the two exercise programs. Interpret the interval in the context of the research.

11.74 An article titled **"TCU Horned Frogs Game Preview (Part 1)" (uwdawgpound.com, 2016)** previews a college basketball game between the University of Washington Huskies and the TCU Horned Frogs. The profile for TCU player Kenrich Williams states, "He has the statistical oddity of shooting better from behind the arc than from the free throw line but that's a marker of small sample sizes." (Note that "behind the arc" means three-point shots, which are taken behind an arc relatively far from the basket.)

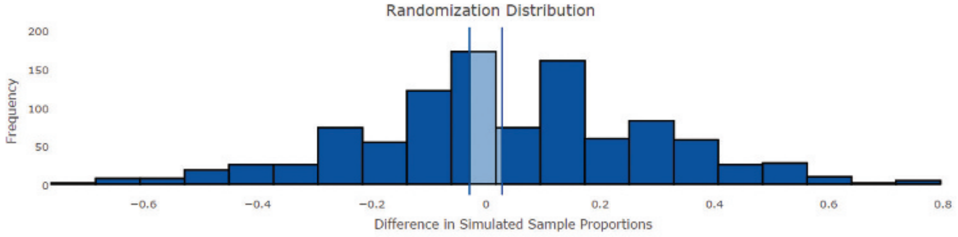
At that point in the season, TCU had played five games. Kenrich Williams had made 4 of the 10 free throws that he had attempted (40%), and 3 of the 7 three-point shots that he had attempted (42.9%). Suppose that it is reasonable to consider these shots to be representative sample of Williams's abilities to make free throws and three-point shots for the 2016–2017 college basketball season.

a. Explain why the data in this exercise should not be analyzed using a large-sample hypothesis test for the difference in two population proportions.

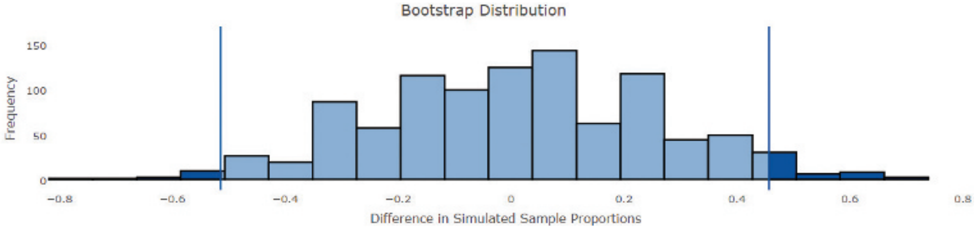
b. Use the output at the bottom of the page from the Shiny app "Randomization Test for Two Proportions" to carry out a hypothesis test to determine if there is evidence that supports the statement that Williams is a better shooter from behind the arc than from the free throw line.

c. Use the output to identify a 95% confidence interval for the difference in the proportions of made free throws and made three-point shots by Kenrich Williams for the 2016–2017 season. Interpret the confidence interval in context.

Output for Exercise 11.74



Observed Difference in Sample Proportions	Standard Error	P-Value	# of Simulated Samples	# of Values in Tail
-0.029	0.247	0.994	1000	994



Observed Difference in Sample Proportions	Standard Error	Lower Bound	Upper Bound	Conf Level	# of Bootstrap Samples
-0.029	0.244	-0.514	0.457	95%	1000

11.75 The report **"Americans' Views on Mobile Etiquette" (pewinternet.org, August 26, 2015)** describes a study in which 354 of 708 people in a sample of 18- to 29-year-olds and 412 of the 1029 people in a sample of 30- to 49-year-olds said that they thought it was OK to use a cell phone in a restaurant. Notice that the sample sizes for the two groups—people age 18 to 29 and those age 30 to 39—are large enough to satisfy the conditions for a large-sample test and a large-sample confidence interval for the difference in two population proportions. Even though the sample sizes are large enough, simulation-based methods can be used.

a. Use the output at the bottom of the page from the Shiny app "Randomization Test for Two Proportions," to carry out a randomization test to determine if there is convincing evidence that the proportion who think it is OK to use a cell phone at a restaurant is higher for the 18 to 29 age group than for the 30 to 49 age group.

b. Use the output from the Shiny app "Bootstrap Confidence Interval for Difference in Two Proportions" to identify a 90% bootstrap confidence interval for the difference in the population proportions of people age 18 to 29 and those age 30 to 39 who would say that they think that it is acceptable to use a cell phone in a restaurant.

c. Use the large-sample z confidence interval to estimate the difference in the population proportions of people age 18 to 29 and those age 30 to 39 who would say that it is acceptable to use a cell phone in a restaurant. Would the interpretation change using the bootstrap confidence interval compared with the large-sample confidence interval? Explain.

11.76 A report in *USA TODAY* described an experiment to explore the accuracy of wearable devices designed to measure heart rate **("Wearable Health Monitors Not Always Reliable, Study Shows," *USA TODAY*, October 12, 2016)**.

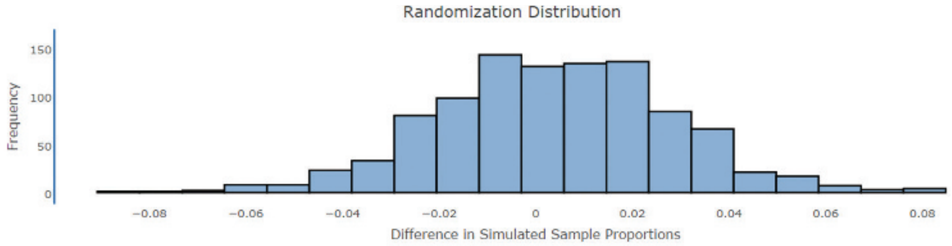
The researchers found that when 50 volunteers wore an Apple Watch to track heart rate as they walked, jogged, and ran quickly on a treadmill for 3 minutes, the results were accurate compared with an EKG 91% of the time. When 50 volunteers wore a Fitbit Charge, the heart rate results were accurate 84% of the time.

a. Explain why the data from this study should not be analyzed using a large-sample hypothesis test for a difference in two population proportions.

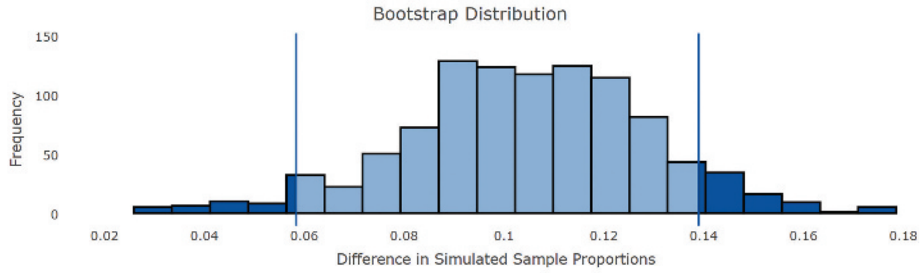
b. Carry out a hypothesis test to determine if there is convincing evidence that the proportion of accurate results for people wearing an Apple Watch is greater than this proportion for those wearing a Fitbit Charge. Use the Shiny app "Randomization Test for Two Proportions" (in the collection at **www.cengage.com/Peck)** to report an approximate P-value and use it to reach a decision in the hypothesis test. Remember to interpret the results of the test in context.

c. Use the Shiny app "Bootstrap Confidence Interval for Difference in Two Proportions" (in the collection at **www.cengage.com/Peck)** to obtain a 95% bootstrap confidence interval for the difference in the population proportions of accurate results for people wearing an Apple Watch and those wearing a Fitbit Charge. Interpret the interval in the context of the research.

Output for Exercise 11.75



Observed Difference in Sample Proportions	Standard Error	P-Value	# of Simulated Samples	# of Values in Tail
0.100	0.024	0.000	1000	0



Observed Difference in Sample Proportions	Standard Error	Lower Bound	Upper Bound	Conf Level	# of Bootstrap Samples
0.100	0.024	0.058	0.139	90%	1000

Copyright 2025 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s). Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

11.77 People in a sample of 258 cell phone users age 20 to 39 were asked if they use their cell phones to stay connected while they are in bed, and 168 said "yes." The same question was also asked of each person in a sample of 129 cell phone users age 40 to 49, and 61 said "yes." You might expect the proportion who stay connected while in bed to be higher for the 20 to 39 age group than for the 40 to 49 age group, but how much higher?

a. Construct and interpret a 90% large-sample confidence interval for the difference in the population proportions of cell phone users age 20 to 39 and those age 40 to 49 who say that they sleep with their cell phones. Interpret the confidence interval in context.

b. Note that the sample sizes in the two groups—cell phone users age 20 to 39 and those age 40 to 49—are large enough to satisfy the conditions for a large-sample test and a large-sample confidence interval for two population proportions. Even though the sample sizes are large enough, simulation-based methods can still be used. Use the output at the bottom of the page to identify a 90% bootstrap confidence interval for the difference in the population proportions of cell phone users age 20 to 39 and those age 40 to 49 who say that they sleep with their cell phones.

c. Compare the confidence intervals calculated in Parts (a) and (b). Would the interpretation change using the bootstrap confidence interval compared with the large-sample confidence interval? Explain.

11.78 The article **"Americans Say No to Electric Cars Despite Gas Prices" (*USA TODAY*, May 25, 2011)** describes a survey of public opinion on issues related to rising gas prices. The survey was conducted by Gallup, a national polling organization. Each person in a representative sample of adult Americans with an annual income less than $30,000 and each person in an independently selected representative sample of adult Americans with an annual income greater than $75,000 was asked whether they would consider buying an electric car if gas prices continue to rise. In the sample of Americans with annual incomes less than $30,000, 65% said that they would not buy an electric car no matter how high gas prices were to rise. In the sample of Americans with annual incomes greater than $75,000, 59% responded this way. The article did not give the sample sizes, but for the purposes of this exercise, suppose the sample sizes were both 300. One question of interest is whether the proportion who would never consider buying an electric car is different for the two income groups.

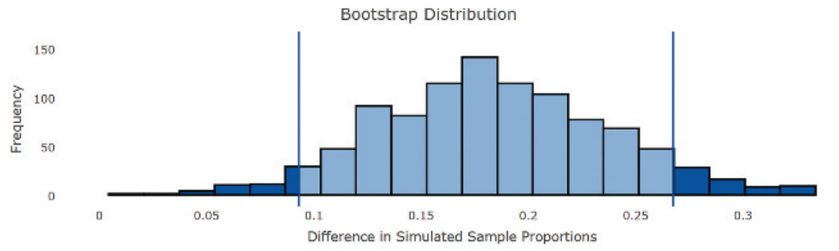
Note that the sample sizes in the two groups are large enough to satisfy the conditions for a large-sample test and a large-sample confidence interval for two population proportions. Even though the sample sizes are large enough, you can still use simulation-based methods.

a. Based on these data, is it reasonable to conclude that the proportions of Americans with an annual income less than $30,000 and Americans with an annual income greater than $75,000 who would never consider buying an electric car differ? Use large-sample methods to test the appropriate hypotheses using a significance level of 0.05.

b. Use the output at the bottom of the page to carry out a randomization test of the same hypotheses tested in Part (a).

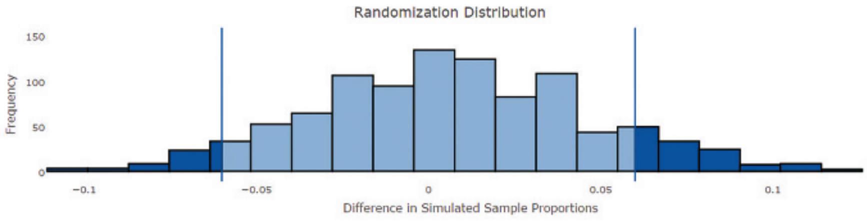
c. Compare the conclusions in Parts (a) and (b). Would the same conclusion be reached in either case? Explain.

Output for Exercise 11.77



Observed Difference in Sample Proportions	Standard Error	Lower Bound	Upper Bound	Conf Level	# of Bootstrap Samples
0.178	0.053	0.093	0.267	90%	1000

Output for Exercise 11.78



Observed Difference in Sample Proportions	Standard Error	P-Value	# of Simulated Samples	# of Values in Tail
0.060	0.040	0.148	1000	148

Copyright 2025 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s). Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

Chapter Activities

● Data set available online

Activity 11.1 Helium-Filled Footballs?

Technology activity: Requires Internet access.

The purpose of this activity is to use sample data to carry out a hypothesis test to determine if helium-filled footballs travel farther than those filled with air.

Background: Do you think that a football filled with helium will travel farther than a football filled with air? Two researchers at the Ohio State University investigated this question by performing an experiment in which 39 people each kicked a helium-filled football and an air-filled football. Half were assigned to kick the air-filled football first and then the helium-filled ball, whereas the other half kicked the helium-filled ball first followed by the air-filled ball. Distance (in yards) was measured for each kick.

In this activity, you will use data from this experiment (from the **Data and Story Library, lib.stat.cmu.edu/DASL**) and then carry out a hypothesis test to determine whether the mean distance is greater for helium-filled footballs than for air-filled footballs.

1. Do you think that helium-filled balls will tend to travel farther than air-filled balls when kicked? Before looking at the data, write a few sentences indicating what you think the outcome of this experiment was and describing the reasoning that supports your prediction.
2. ● Download the data set for Activity 11.1 from the student resource website or the WebAssign resources tab.
3. There are two samples in this data set. One consists of distances traveled for the 39 kicks of the air-filled football, and the other consists of the 39 distances for the helium-filled football. Are these samples independent or paired? Explain.
4. Carry out an appropriate hypothesis test to determine whether there is convincing evidence that the mean distance traveled is greater for a helium-filled football than for an air-filled football.
5. Is the conclusion in the test of Step 4 consistent with your initial prediction of the outcome of this experiment? Explain.
6. Write a paragraph for the sports section of your school newspaper describing this experiment and the conclusions that can be drawn from it.

Activity 11.2 Thinking About Data Collection

The purpose of this activity is to allow you to demonstrate your understanding of the difference between paired and independent samples.

Background: In this activity you will design two experiments that would allow you to investigate whether people tend to have quicker reflexes when reacting with their dominant hand than with their nondominant hand.

1. Working in a group, design an experiment to investigate the given research question that would result in independent samples. Be sure to describe how you plan to measure quickness of reflexes, what extraneous variables will be directly controlled, and the role that randomization plays in your design.
2. How would you modify the design from Step 1 so that the resulting data are paired? Is the way in which randomization is incorporated into the new design different from the way it is incorporated in the design from Step 1? Explain.
3. Which of the two proposed designs would you recommend, and why?
4. If assigned to do so by your instructor, carry out one of your experiments and analyze the resulting data. Write a brief report that describes the experimental design, includes both graphical and numerical summaries of the resulting data, and communicates the conclusions that follow from your data analysis.

Activity 11.3 A Meaningful Paragraph

The purpose of this activity is to allow you to demonstrate your understanding of the terminology used in hypothesis testing.

Write a meaningful paragraph that includes the following six terms: **paired samples, significantly different, *P*-value, sample, population, alternative hypothesis.**

A "meaningful paragraph" is a coherent piece of writing in an appropriate context that uses all of the listed words. The paragraph should show that you understand the meaning of the terms and their relationship to one another. A sequence of sentences that just define the terms is *not* a meaningful paragraph. When choosing a context, think carefully about the terms you need to use. Choosing a good context will make writing a meaningful paragraph easier.

Copyright 2025 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s). Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

Summary Key Concepts and Formulas

Term or Formula	Comment
Independent samples	Two samples where the individuals or objects in the first sample are selected independently from those in the second sample.
Paired samples	Two samples for which each observation in one sample is paired in a meaningful way with a particular observation in a second sample.
$t = \dfrac{(\bar{x}_1 - \bar{x}_2) - \text{hypothesized value}}{\sqrt{\dfrac{s_1^2}{n_1} + \dfrac{s_2^2}{n_2}}}$	The test statistic for testing H_0: $\mu_1 - \mu_2 =$ hypothesized value when the samples are independently selected and the sample sizes are large or it is reasonable to assume that both population distributions are normal.
$(\bar{x}_1 - \bar{x}_2) \pm (t \text{ critical value})\sqrt{\dfrac{s_1^2}{n_1} + \dfrac{s_2^2}{n_2}}$	A formula for constructing a confidence interval for $\mu_1 - \mu_2$ when the samples are independently selected and the sample sizes are large or it is reasonable to assume that the population distributions are normal.
$\text{df} = \dfrac{(V_1 + V_2)^2}{\dfrac{V_1^2}{n_1 - 1} + \dfrac{V_2^2}{n_2 - 1}}$ where $V_1 = \dfrac{s_1^2}{n_1}$ and $V_2 = \dfrac{s_2^2}{n_2}$	The formula for determining df for the two-sample t test and confidence interval.
$\bar{x}_d$	The sample mean difference.
s_d	The standard deviation of the sample differences.
μ_d	The mean value for the population of differences.
σ_d	The standard deviation for the population of differences.
$t = \dfrac{\bar{x}_d - \text{hypothesized value}}{\dfrac{s_d}{\sqrt{n}}}$	The paired t test statistic for testing H_0: $\mu_d =$ hypothesized value.
$\bar{x}_d \pm (t \text{ critical value})\dfrac{s_d}{\sqrt{n}}$	The paired t confidence interval formula.
$\hat{p}_c = \dfrac{n_1\hat{p}_1 + n_2\hat{p}_2}{n_1 + n_2}$	$\hat{p}_c$ is the statistic for estimating the common population proportion when $p_1 = p_2$.
$z = \dfrac{\hat{p}_1 - \hat{p}_2}{\sqrt{\dfrac{\hat{p}_c(1 - \hat{p}_c)}{n_1} + \dfrac{\hat{p}_c(1 - \hat{p}_c)}{n_2}}}$	The test statistic for testing H_0: $p_1 - p_2 = 0$ when the samples are independently selected and both sample sizes are large.
$(\hat{p}_1 - \hat{p}_2) \pm (z \text{ critical value})\sqrt{\dfrac{\hat{p}_1(1 - \hat{p}_1)}{n_1} + \dfrac{\hat{p}_2(1 - \hat{p}_2)}{n_2}}$	A formula for constructing a confidence interval for $p_1 - p_2$ when both sample sizes are large.

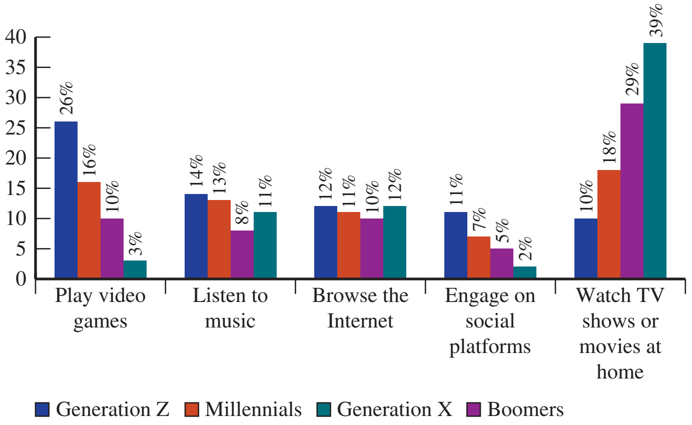
Chapter Review Exercises 11.79–11.91

● Data set available online

11.79 The authors of the paper **"'It's Not Rocket Science' and 'It's Not Brain Surgery'—It's a Walk in the Park": Prospective Comparative Study" (*British Medical Journal* [2021]: e067883)** wondered if brain surgeons or rocket scientists were better problem solvers. Each person in a sample of aerospace engineers and a sample of neurosurgeons was scored on a spatial problem-solving task. They reported a 95% confidence interval for the difference in mean scores (neurosurgeons – aerospace engineers) of (−0.39, 0.01). Provide an interpretation of this interval in the context of this problem.

11.80 The figure at the top of the next page is similar to one that appeared in the report **"Digital Media Trends, 15th Edition: Courting the Consumer in a World of Choice" (deloitte.com, April 16, 2021)**. The graph summarizes responses by generation to the question: What is your favorite entertainment activity? Suppose that the percentages given in the graph are based on

Generation Z may not be "video first"
Respondents across generations ranked their favorite entertainment activity



independent representative samples of 500 people from each of the four generations. Is there convincing evidence that the proportion of Boomers who say that their favorite entertainment activity is playing video games is less than this proportion for Generation X? Test the relevant hypotheses using $\alpha = 0.05$.

11.81 Refer back to the information given in the previous exercise.

a. Using the percentages in the figure for the previous exercise and assuming sample sizes of 500 for each generation, carry out a hypothesis test to determine if there is convincing evidence that there is a difference between the proportion of Boomers and the proportion of Generation X who chose browsing the Internet as their favorite entertainment activity. Use a significance level of 0.05 for your test.

b. Suppose that the percentages given in the figure were based on samples of 2000 from each generation. Is there convincing evidence that there is a difference between the proportion of Boomers and the proportion of Generation X who chose browsing the Internet as their favorite entertainment activity. Use a significance level of 0.05 for your test.

c. Explain why the hypothesis tests in Parts (a) and (b) resulted in different conclusions.

11.82 When a surgeon repairs injuries, sutures (stitched knots) are used to hold together and stabilize the injured area. If these knots elongate and loosen through use, the injury may not heal properly because the tissues would not be optimally positioned. Researchers at the University of California, San Francisco, tied a series of different types of knots with two types of suture material, Maxon and Ticron.

Suppose that 112 tissue specimens were available and that for each specimen the type of knot and suture material were randomly assigned. The investigators tested the knots to see how much the loops elongated. The elongations (in mm) were measured and the resulting data are summarized here.
For purposes of this exercise, assume it is reasonable to regard the elongation distributions as approximately normal.

	Maxon		
Types of Knot	n	$\bar{x}$	**sd**
Square (control)	10	10.0	0.1
Duncan Loop	15	11.0	0.3
Overhand	15	11.0	0.9
Roeder	10	13.5	0.1
Snyder	10	13.5	2.0

	Ticron		
Types of Knot	n	$\bar{x}$	**sd**
Square (control)	10	2.5	0.06
Duncan Loop	11	10.9	0.40
Overhand	11	8.1	1.00
Roeder	10	5.0	0.04
Snyder	10	8.1	0.06

a. Is there a significant difference in mean elongation between the square knot and the Duncan loop for Maxon thread?

b. Is there a significant difference in mean elongation between the square knot and the Duncan loop for Ticron thread?

c. For the Duncan loop knot, is there a significant difference in mean elongation between the Maxon and Ticron threads?

11.83 The paper **"Beyond the Beauty of Occlusion: Medical Masks Increase Facial Attractiveness More Than Other Face Coverings" (*Cognitive Research: Principles and Implications* [2022])** describes an experiment to investigate if peoples' perception of attractiveness is affected by wearing facial coverings of various types. One of the types of facial coverings considered was a disposable medical mask. Suppose that an experiment was carried out in which participants rated the attractiveness of a person in a photo. Some people saw the person without a mask and others saw a photo of the same person wearing a medical mask. For purposes of this exercise, suppose that 60 people were randomly assigned to one of two groups, with 30 in each group. Participants in the no-mask group rated the attractiveness of a person who was not wearing a mask on a scale of 1 to 7. Participants in the mask group rated the attractiveness of the same person wearing a medical mask. The accompanying summary statistics were approximated from a graph that appeared in the paper.

Copyright 2025 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s). Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

Experimental Group	n	$\bar{x}$	s
Viewed photo of person with no mask	30	3.4	0.14
Viewed photo of person with medical mask	30	3.7	0.15

a. If you wanted to carry out a test to determine if there is convincing evidence that the mean attractiveness rating is not the same for the two experimental groups, would you use a paired-samples *t* test or a two-sample *t* test?

b. Carry out a test to determine if there is convincing evidence that the mean attractiveness rating is not the same for the two experimental groups. Use a significance level of $\alpha = 0.05$ for your test.

c. The authors of the paper concluded that faces were considered more attractive when covered with a medical mask. Is this consistent with the result of your test from Part (b)? Explain.

11.84 ● In a study of memory recall, eight students from a large psychology class were selected at random and given 10 minutes to memorize a list of 20 nonsense words. Each was asked to list as many of the words as they could remember both 1 hour and 24 hours later. The data are as shown in the accompanying table.

Is there evidence to suggest that the mean number of words recalled after 1 hour is greater than the mean recall after 24 hours by more than 3? Use a level 0.01 test.

Subject	1	2	3	4	5	6	7	8
1 hour later	14	12	18	7	11	9	16	15
24 hours later	10	4	14	6	9	6	12	12

11.85 In a test of hypotheses about a difference in treatment proportions, what does it mean when the null hypothesis is not rejected?

11.86 Many fundraisers ask for donations using email and text messages. The paper **"Now or Never! The Effect of Deadlines on Charitable Giving: Evidence from Two Natural Field Experiments" (*Journal of Behavioral and Experimental Economics* [2016]: 1–10)** describes an experiment to investigate whether the proportion of people who make a donation when asked for a donation by email is different from the proportion of people who make a donation when asked for a donation in a text message. In this experiment, 1.32% of those who received and opened an email request for a donation and 7.77% of those who received a text message asking for a donation actually made a donation. Assume that the people who received these requests were randomly assigned to one of the two groups (email or text message) and suppose that the given percentages are based on sample sizes of 2000 (the actual sample sizes in the experiment were much larger).

a. The study described is an experiment with two treatments. What are the two treatments?

b. Is there convincing evidence that the proportion who make a donation is not the same for the two different methods? Carry out a hypothesis test using a significance level of 0.05.

c. Use a 90% confidence interval to estimate the difference in the proportions who donate for the two different treatments.

11.87 Data from a survey of teens and parents of teens are summarized in the report **"How Teens and Parents Navigate Screen Time and Device Distractions" (Pew Research Center, August 22, 2018).** Participants in independent representative samples of teens and parents of teens were asked if they think that they spend too much time on their cell phones. Of the 743 teens surveyed, 54% said "yes" and of the 1058 parents of teens surveyed, 36% said "yes."

a. Construct and interpret a 95% large-sample confidence interval for the difference in the proportion of teens who think they spend too much time on their cell phones and this proportion for parents of teens.

b. Is 0 included in the confidence interval of Part (a)? What does this imply about the difference in population proportions?

11.88 The survey referenced in the previous exercise also asked participants if they check their phone for messages as soon as they wake up. Of the 743 teens surveyed, 44% said "yes" and of the 1058 parents of teens surveyed, 26% said "yes." Do the data provide convincing evidence that the proportion of parents of teens who check their phone for messages as soon as they wake up is less than the proportion of teens who check for messages as soon as they wake up? Test the relevant hypotheses using a significance level of 0.05.

11.89 Here's one to sink your teeth into: The authors of the article **"Analysis of Food Crushing Sounds During Mastication: Total Sound Level Studies" (*Journal of Texture Studies* [1990]: 165–178)** studied the nature of sounds generated during eating. Peak loudness (in decibels at 20 cm away) was measured for both open-mouth and closed-mouth chewing of potato chips and of tortilla chips. Forty subjects participated, with 10 assigned at random to each combination of conditions (such as closed-mouth potato chip, and so on). We are not making this up! Summary values taken from plots given in the

Copyright 2025 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s). Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

article appear in the accompanying table. For purposes of this exercise, suppose that it is reasonable to regard the peak loudness distributions as approximately normal.

	n	$\bar{x}$	s
Potato Chip			
Open mouth	10	63	13
Closed mouth	10	54	16
Tortilla Chip			
Open mouth	10	60	15
Closed mouth	10	53	16

a. Construct a 95% confidence interval for the difference in mean peak loudness between open-mouth and closed-mouth chewing of potato chips. Interpret the resulting interval.

b. For closed-mouth chewing (the recommended method!), is there sufficient evidence to indicate that there is a difference between potato chips and tortilla chips with respect to mean peak loudness? Test the relevant hypotheses using $\alpha = 0.01$.

c. The means and standard deviations given here were actually for stale chips. When ten measurements of peak loudness were recorded for closed-mouth chewing of fresh tortilla chips, the resulting mean and standard deviation were 56 and 14, respectively. Is there sufficient evidence to conclude that chewing fresh tortilla chips is louder than chewing stale chips? Use $\alpha = 0.05$.

11.90 Descriptions of three studies are given. In each of the studies, the two populations of interest are students majoring in science at a particular university and students majoring in liberal arts at this university. For each of these studies, indicate whether the samples are independently selected or paired.

Study 1: To determine if there is evidence that the mean number of hours spent studying per week differs for the two populations, a random sample of 100 science majors and a random sample of 75 liberal arts majors are selected.

Study 2: To determine if the mean amount of money spent on textbooks differs for the two populations, a random sample of science majors is selected. Each student in this sample is asked how many units they are enrolled in for the current semester. For each of these science majors, a liberal arts major who is taking the same number of units is identified and included in the sample of liberal arts majors.

Study 3: To determine if the mean amount of time spent using the campus library differs for the two populations, a random sample of science majors is selected. A separate random sample of the same size is selected from the population of liberal arts majors.

11.91 ● Both surface soil and subsoil specimens were taken from eight randomly selected agricultural locations in a particular county. The soil specimens were analyzed to determine both surface pH and subsoil pH, with the results shown in the accompanying table.

Location	1	2	3	4	5	6	7	8
Surface pH	6.55	5.98	5.59	6.17	5.92	6.18	6.43	5.68
Subsoil pH	6.78	6.14	5.80	5.91	6.10	6.01	6.18	5.88

a. Calculate a 90% confidence interval for the mean difference between surface and subsoil pH for agricultural land in this county.

b. What assumptions are necessary for the interval in Part (a) to be valid?

Technology Notes

Two-sample *t* Test for $\mu_1 - \mu_2$

JMP

1. Input the raw data for both groups into the first column
2. Input the group information into the second column
3. Click **Analyze** and select **Fit Y by X**
4. Click and drag the first column's name from the box under **Select Columns** to the box next to **Y, Response**
5. Click and drag the second column's name from the box under **Select Columns** to the box next to **X, Factor**
6. Click **OK**
7. Click the red arrow next to **Oneway Analysis of…** and select **t Test**

Minitab

Summarized data

1. Click **Stat** then click **Basic Statistics** then click **2-sample t…**
2. Select **Summarized data** from the dropdown menu
3. Summary statistics for the first sample will be entered in the **Sample 1** column. Type the value for the sample size, in the **Sample size:** row, type the value for the sample mean in the **Sample mean:** row, and type the value for the sample standard deviation in the **Standard deviation:** row.
4. Summary statistics for the second sample will be entered in the **Sample 2** column. Type the value for the sample size, in the **Sample size:** row, type the value for the sample mean

Copyright 2025 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s). Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

in the **Sample mean:** row, and type the value for the sample standard deviation in the **Standard deviation:** row.

5. Click **Options…**
6. Input the appropriate hypothesized value for the difference of the population means in the box next to **Hypothesized difference:**
7. Select the appropriate alternative from the drop-down menu next to **Alternative hypothesis:**
8. Click **OK**
9. Click **OK**

Note: You may also run this test with the assumption of equal variances by clicking the checkbox next to **Assume equal variances** after Step 8 in the above sequence.

Raw data

1. Input the raw data into two separate columns
2. Click **Stat** then click **Basic Statistics** then click **2-sample t…**
3. Select **Each sample is in its own column** from the drop-down menu
4. Click in the box next to **Sample 1:**
5. Double-click the column name where the first group's data is stored
6. Click in the box next to **Sample 2:**
7. Double-click the column name where the second group's data is stored
8. Click **Options…**
9. Input the appropriate hypothesized value for the difference of the population means in the box next to **Hypothesized difference:**
10. Select the appropriate alternative from the drop-down menu next to **Alternative hypothesis:**
11. Click **OK**
12. Click **OK**

Note: You may also run this test with the assumption of equal variances by clicking the checkbox next to **Assume equal variances** after Step 11 in the above sequence.

SPSS

1. Input the raw data for BOTH groups into the first column
2. Input the data for groups into the second column (input A for the first group and B for the second group)
3. Click **Analyze** then click **Compare Means** then click **Independent-Samples T Test**
4. Click the name of the column containing the raw data and click the arrow to move this variable to the **Test Variable(s):** box
5. Click the name of the column containing the group information and click the arrow to move this variable to the **Grouping Variable:** box
6. Click the **Define Groups…** button
7. In the box next to **Group 1:** type A
8. In the box next to **Group 2:** type B
9. Click **Continue**
10. Click **OK**

Note: This procedure produces two-sample *t* tests under the assumption of equal variances AND also when equal variances are not assumed. It also outputs a two-tailed *P*-value.

Excel

1. Input the raw data for each group into two separate columns
2. Click on the **Data** ribbon
3. Click **Data Analysis** in the **Analysis** group

Note: If you do not see **Data Analysis** listed on the Ribbon, see the Technology Notes for Chapter 2 for instructions on installing this add-on.

If you are performing a test where you are assuming equal variances, continue with Steps 4–9 below. If you are performing a test where you are not assuming equation variances, skip to Step 10.

4. Select **t-Test: Two Samples Assuming Equal Variances** from the dialog box and click **OK**
5. Click in the box next to **Variable 1 Range:** and select the first column of data
6. Click in the box next to **Variable 2 Range** and select the second column of data (if you have used and selected column titles for BOTH variables, select the check box next to **Labels**)
7. Click in the box next to **Hypothesized Mean Difference** and type your hypothesized value (in general, this will be 0)
8. Click in the box next to **Alpha:** and type in the significance level
9. Click **OK**
10. Select **t-Test: Two Samples Assuming Unequal Variances** from the dialog box and click **OK**
11. Click in the box next to **Variable 1 Range:** and select the first column of data
12. Click in the box next to **Variable 2 Range** and select the second column of data (if you have used and selected column titles for BOTH variables, select the check box next to **Labels**)
13. Click in the box next to **Hypothesized Mean Difference** and type your hypothesized value (in general, this will be 0)
14. Click in the box next to **Alpha:** and type in the significance level
15. Click **OK**

Note: This procedure outputs *P*-values for both a one-sided and two-sided test.

TI-83/84

Summarized data

1. Press **STAT**
2. Highlight **TESTS**
3. Highlight **2-SampT-Test…**
4. Highlight **Stats** and press **ENTER**
5. Next to $\bar{x}1$ input the value for the sample mean from the first sample
6. Next to **Sx1** input the value for the sample standard deviation from the first sample
7. Next to **n1** input the value for the sample size from the first sample
8. Next to $\bar{x}2$ input the value for the sample mean from the second sample
9. Next to **Sx2** input the value for the sample standard deviation from the second sample

10. Next to **n2** input the value for the sample size from the second sample
11. Next to μ**1** highlight the appropriate alternative hypothesis and press **ENTER**
12. Highlight **Calculate** and press **ENTER**

Note: You may also run this test with the assumption of equal variances by selecting Yes next to **Pooled:**

Raw data

1. Input raw data for the first group into **L1** and input raw data for the second group into **L2** (In order to access lists press the **STAT** key, highlight the option called **Edit…** then press **ENTER**)
2. Press **STAT**
3. Highlight **TESTS**
4. Highlight **2-SampT-Test…**
5. Highlight **Data** and press **ENTER**
6. Next to μ**1** highlight the appropriate alternative hypothesis and press **ENTER**
7. Highlight **Calculate** and press **ENTER**

Note: You may also run this test with the assumption of equal variances by selecting Yes next to **Pooled:**

TI-Nspire

Summarized data

1. Under **Scratchpad** select **A Calculate**
2. Press the **menu** key and select **6:Statistics** then **7:Stat Tests** then **4:2-Sample t test…** and press **enter**
3. From the drop-down menu select **Stats**
4. Press **OK**
5. Next to $\bar{x}1$ input the value for the sample mean for the first sample
6. Next to **Sx1** input the value for the sample standard deviation for the first sample
7. Next to **n1** input the value for the sample size for the first sample
8. Next to $\bar{x}2$ input the value for the sample mean for the second sample
9. Next to **Sx2** input the value for the sample standard deviation for the second sample
10. Next to **n2** input the value for the sample size for the second sample
11. Next to **Alternate Hyp** select the appropriate alternative hypothesis from the drop-down menu
12. Press **OK**

Note: You may also run this test with the assumption of equal variances by selecting Yes in the drop-down menu next to **Pooled:**

Raw data

1. Enter the raw data for each group into separate data lists (In order to access data lists select the spreadsheet option and press **enter**)

   Note: Be sure to title the list by selecting the top row of the column and typing a title.

2. Press the **menu** key and select **4:Statistics** then **4:Stat Tests** then **4:2-Sample t test…** and press **enter**
3. From the drop-down menu select **Data**
4. Press **OK**
5. Next to **List 1** select the list containing your data from the first sample
6. Next to **List 2** select the list containing your data from the second sample
7. Next to **Alternate Hyp** select the appropriate alternative hypothesis from the drop-down menu
8. Press **OK**

Note: You may also run this test with the assumption of equal variances by selecting Yes in the drop-down menu next to **Pooled:**

Paired *t* Test for Difference of Population Means

JMP

1. Enter the data for one sample in the first column
2. Enter the paired data from the second sample in the second column
3. Click **Analyze**, then **Specialized Modeling**, then select **Matched Pairs**
4. Click and drag the first column name from the box under **Select Columns** to the box next to **Y, Paired Response**
5. Click and drag the second column name from the box under **Select Columns** to the box next to **Y, Paired Response**
6. Click **OK**

Minitab

Summarized data

1. Click **Stat** then click **Basic Statistics** then click **Paired t…**
2. Select **Summarized data (differences)** from the dropdown menu
3. In the box next to **Sample size:** type the sample size, *n*
4. In the box next to **Sample mean:** type the sample mean for the DIFFERENCE of each pair of data values
5. In the box next to **Standard deviation:** type the sample standard deviation for the DIFFERENCE of each pair of data values
6. Click **Options…**
7. Input the appropriate hypothesized value for the difference of the paired population means in the box next to **Hypothesized difference**
8. Select the appropriate alternative from the drop-down menu next to **Alternative hypothesis:**
9. Click **OK**
10. Click **OK**

Raw data

1. Input the raw data into two separate columns
2. Click **Stat** then click **Basic Statistics** then click **Paired t…**
3. Select **Each sample is in a column** from the dropdown menu
4. Click in the box next to **Sample 1:**
5. Double click the column name where the first group's data are stored
6. Click in the box next to **Sample 2:**
7. Double click the column name where the second group's data are stored

Copyright 2025 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s). Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

8. Click **Options...**
9. Input the appropriate hypothesized value for the difference of the paired population means in the box next to **Hypothesized difference:**
10. Select the appropriate alternative hypothesis from the drop-down menu next to **Alternative hypothesis:**
11. Click **OK**
12. Click **OK**

SPSS

1. Input the data for each group into two separate columns
2. Click **Analyze** then click **Compare Means** then click **Paired-Samples T Test**
3. Select the first column and click the arrow to move it into the Pair 1 row, Variable 1 column
4. Select the second column and click the arrow to move it to the Pair 1 row, Variable 2 column
5. Click **OK**

Note: This procedure produces a two-sided P-value.

Excel

1. Input the raw data for each group into two separate columns
2. Click on the **Data** ribbon
3. Click **Data Analysis** in the **Analysis** group

Note: If you do not see **Data Analysis** listed on the Ribbon, see the Technology Notes for Chapter 2 for instructions on installing this add-on.

4. Select **t-Test: Paired Two Sample for Means** from the dialog box and click **OK**
5. Click in the box next to **Variable 1 Range:** and select the first column of data
6. Click in the box next to **Variable 2 Range** and select the second column of data (if you have used and selected column titles for BOTH variables, select the check box next to **Labels**)
7. Click in the box next to **Hypothesized Mean Difference** and type your hypothesized value (in general, this will be 0)
8. Click in the box next to **Alpha:** and type in the significance level
9. Click **OK**

Note: This procedure outputs P-values for both a one-sided and two-sided test.

TI-83/84

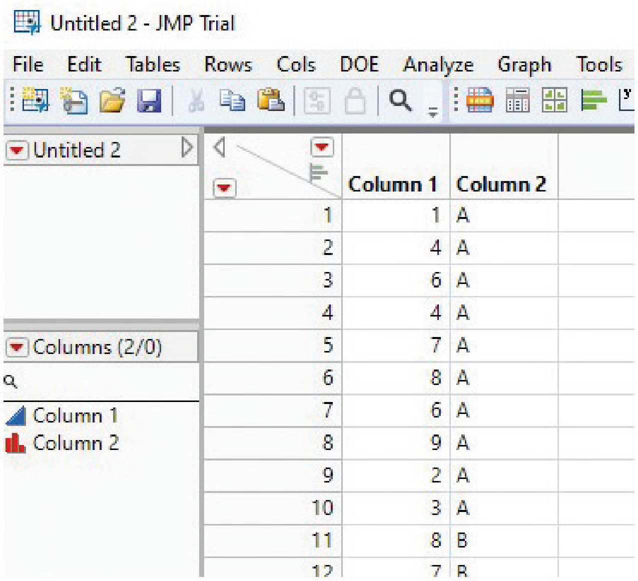
The TI-83/84 does not provide the option for a paired t test. However, one can be found by entering the difference data into a list and following the t Test for Population Mean, μ procedure in Chapter 10.

TI-Nspire

The TI-Nspire does not provide the option for a paired t test. However, one can be found by entering the difference data into a list and following the t Test for Population Mean, μ procedure in Chapter 10.

Confidence Interval for $\mu_1 - \mu_2$

1. Input the raw data for both groups into the first column
2. Input the group information into the second column



	Column 1	Column 2
1	1	A
2	4	A
3	6	A
4	4	A
5	7	A
6	8	A
7	6	A
8	9	A
9	2	A
10	3	A
11	8	B
12	7	B

Source: JMP

3. Click **Analyze** and select **Fit Y by X**
4. Click and drag the first column's name from the box under **Select Columns** to the box next to **Y, Response**
5. Click and drag the second column's name from the box under **Select Columns** to the box next to **X, Factor**
6. Click **OK**
7. Click the red arrow next to **Oneway Analysis of…** and select **t Test**

Note: The 95% confidence interval is automatically displayed. To change the confidence level, click the red arrow next to **Oneway Analysis of…** and select **Set α Level** then click the appropriate α-level or select **Other** and type the appropriate level.

Minitab

Summarized data

1. Click **Stat** then click **Basic Statistics** then click **2-sample t…**
2. Select **Summarized data** from the dropdown menu
3. Summary statistics for the first sample will be entered in the **Sample 1** column. Type the value for the sample size, in the **Sample size:** row, type the value for the sample mean in the **Sample mean:** row, and type the value for the sample standard deviation in the **Standard deviation:** row.
4. Summary statistics for the second sample will be entered in the **Sample 2** column. Type the value for the sample size, in the **Sample size:** row, type the value for the sample mean in the **Sample mean:** row, and type the value for the sample standard deviation in the **Standard deviation:** row.
5. Click **Options…**
6. Input the appropriate confidence level in the box next to **Confidence Level**
7. Click **OK**
8. Click **OK**

Raw data

1. Input the raw data for each group into a separate column
2. Click **Stat** then click **Basic Statistics** then click **2-sample t…**
3. Select **Each sample is in its own column** from the dropdown menu
4. Click in the box next to **Sample 1:**
5. Double click the column name where the first group's data are stored
6. Click in the box next to **Sample 2:**
7. Double-click the column name where the second group's data are stored
8. Click **Options…**
9. Input the appropriate confidence level in the box next to **Confidence Level**
10. Click **OK**
11. Click **OK**

SPSS

1. Input the raw data for BOTH groups into the first column
2. Input the data for groups into the second column (input A for the first group and B for the second group)
3. Click **Analyze** then click **Compare Means** then click **Independent-Samples T Test**
4. Click the name of the column containing the raw data and click the arrow to move this variable to the **Test Variable(s):** box
5. Click the name of the column containing the group information and click the arrow to move this variable to the **Grouping Variable:** box
6. Click the **Define Groups…** button
7. In the box next to **Group 1:** type A
8. In the box next to **Group 2:** type B
9. Click **Continue**
10. Click **Options…**
11. Input the confidence level in the box next to **Confidence Interval Percentage:**
12. Click **Continue**
13. Click **OK**

Note: This procedure produces confidence intervals under the assumption of equal variances AND also when equal variances are not assumed.

Excel

Excel does not have the functionality to produce a confidence interval automatically for the difference of two population means. However, you can manually type the formulas for the lower and upper limits into two separate cells and have Excel calculate the results for you. You may also use Excel to find the *t* critical value based on the confidence level using the following steps:

1. Click in an empty cell
2. Click **Formulas**
3. Click **Insert Function**
4. Select **Statistical** from the drop-down box for the category
5. Select **TINV** and click **OK**
6. In the box next to **Probability** type in the value representing one minus your selected confidence level
7. In the box next to **Deg_freedom** type in the degrees of freedom
8. Click **OK**

TI-83/84

Summarized data

1. Press **STAT**
2. Highlight **TESTS**
3. Highlight **2-SampTInt…** and press **ENTER**
4. Highlight **Stats** and press **ENTER**
5. Next to $\bar{x}1$ input the value for the sample mean from the first sample
6. Next to **Sx1** input the value for the sample standard deviation from the first sample
7. Next to **n1** input the value for the sample size from the first sample
8. Next to $\bar{x}2$ input the value for the sample mean from the second sample
9. Next to **Sx2** input the value for the sample standard deviation from the second sample
10. Next to **n2** input the value for the sample size from the second sample
11. Next to **C-Level** input the appropriate confidence level
12. Highlight **Calculate** and press **ENTER**

Raw data

1. Enter the data into **L1** and **L2** (In order to access lists press the **STAT** key, then press **ENTER**)
2. Press **STAT**
3. Highlight **TESTS**
4. Highlight **2-SampTInt…** and press **ENTER**
5. Highlight **Data** and press **ENTER**
6. Next to **C-Level** input the appropriate confidence level
7. Highlight **Calculate** and press **ENTER**

TI-Nspire

Summarized data

1. Under **Scratchpad** select **A Calculate**
2. Press the **menu** key and select **6:Statistics** then **6:Confidence Intervals** then **4:2-Sample t Interval…** and press **enter**
3. From the drop-down menu select **Stats**
4. Press **OK**
5. Next to $\bar{x}1$ input the value for the sample mean from the first sample
6. Next to **Sx1** input the value for the sample standard deviation from the first sample
7. Next to **n1** input the value for the sample size from the first sample
8. Next to $\bar{x}2$ input the value for the sample mean from the second sample
9. Next to **Sx2** input the value for the sample standard deviation from the second sample
10. Next to **n2** input the value for the sample size from the second sample
11. Next to **C Level** input the appropriate confidence level
12. Press **OK**

Copyright 2025 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s). Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

Raw data

1. Enter the data into two separate data lists (In order to access data lists select the spreadsheet option and press **enter**)

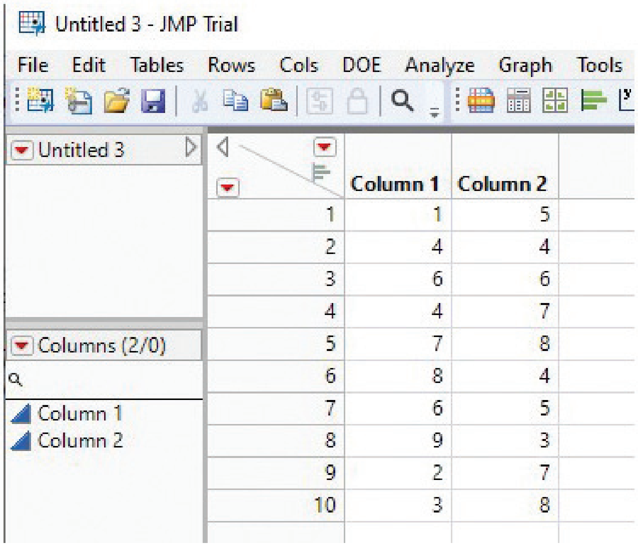
Note: Be sure to title the lists by selecting the top row of the column and typing a title.

2. Press the **menu** key and select **4:Statistics** then **3:Confidence Intervals** then **4: 2-Sample t Interval...** and press **enter**
3. From the drop-down menu select **Data**
4. Press **OK**
5. Next to **List 1** select the list containing the first data sample from the drop-down menu
6. Next to **List 2** select the list containing the second data sample from the drop-down menu
7. Next to **C-Level** input the appropriate confidence level
8. Press **OK**

Confidence Interval for Paired Data

JMP

1. Enter the data for one group in the first column
2. Enter the paired data from the second group in the second column



	Column 1	Column 2
1	1	5
2	4	4
3	6	6
4	4	7
5	7	8
6	8	4
7	6	5
8	9	3
9	2	7
10	3	8

Source: JMP

3. Click **Analyze**, then **Specialized Modeling**, then select **Matched Pairs**
4. Click and drag the first column name from the box under **Select Columns** to the box next to **Y, Paired Response**
5. Click and drag the second column name from the box under **Select Columns** to the box next to **Y, Paired Response**
6. Click **OK**

Note: The 95% confidence interval is automatically displayed. To change the confidence level, click the red arrow next to **Matched Pairs** and select **Set α Level** then click the appropriate a-level or select **Other** and type the appropriate level.

Minitab

Summarized data

1. Click **Stat** then click **Basic Statistics** then click **Paired t…**
2. Select **Summarized data (differences)** from the dropdown menu
3. In the box next to **Sample size:** type the sample size, *n*
4. In the box next to **Sample mean:** type the sample mean for the DIFFERENCE of each pair of data values
5. In the box next to **Standard deviation:** type the sample standard deviation for the DIFFERENCE of each pair of data values
6. Click **Options…**
7. Input the appropriate confidence level in the box next to **Confidence Level**
8. Click **OK**
9. Click **OK**

Raw data

1. Input the raw data into two separate columns
2. Click **Stat** then click **Basic Statistics** then click **Paired t…**
3. Select **Each sample is in a column** from the dropdown menu
4. Click in the box next to **Sample 1:**
5. Double click the column name where the first group's data are stored
6. Click in the box next to **Sample 2:**
7. Double click the column name where the second group's data are stored
8. Click **Options...**
9. Input the appropriate confidence level in the box next to **Confidence level:**
10. Click **OK**
11. Click **OK**

SPSS

1. Input the data for each group into two separate columns
2. Click **Analyze** then click **Compare Means** then click **Paired-Samples T Test**
3. Select the first column and click the arrow to move it into the Pair 1 row, Variable 1 column
4. Select the second column and click the arrow to move it to the Pair 1 row, Variable 2 column
5. Click **Options…**
6. Input the appropriate confidence level in the box next to **Confidence Interval Percentage:**
7. Click **Continue**
8. Click **OK**

Excel

Excel does not have the functionality to produce a confidence interval automatically for paired population means. However, you can manually type the formulas for the lower and upper limits into two separate cells and have Excel calculate the results for you. You may also use Excel to find the *t* critical value based on the confidence level using the following steps:

1. Click in an empty cell
2. Click **Formulas**
3. Click **Insert Function**
4. Select **Statistical** from the drop-down box for the category
5. Select **TINV** and click **OK**
6. In the box next to **Probability** type in the value representing one minus your selected confidence level

Copyright 2025 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s). Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

7. In the box next to **Deg_freedom** type in the degrees of freedom ($n - 1$)
8. Click **OK**

TI-83/84

The TI-83/84 does not provide the option for a paired *t* confidence interval. However, one can be found by entering the difference data into a list and following the Confidence Interval for μ Based on *t*-distribution procedure in Chapter 9.

TI-Nspire

The TI-Nspire does not provide the option for a paired *t* confidence interval. However, one can be found by entering the difference data into a list and following the Confidence Interval for μ Based on *t*-distribution procedure in Chapter 9.

z Test for $p_1 - p_2$

JMP

JMP does not have the functionality to automatically provide the results of a *z* test for the difference of two proportions.

Minitab

Summarized data

1. Click **Stat** then click **Basic Statistics** then click **2 Proportions…**
2. Select **Summarized data** from the dropdown menu
3. Summary statistics for the first sample will be entered in the **Sample 1** column. Type the number of successes in the **Number of events:** row. Type the sample size in the **Number of trials:** row.
4. Summary statistics for the second sample will be entered in the **Sample 2** column. Type the number of successes in the **Number of events:** row. Type the sample size in the **Number of trials:** row.
5. Click **Options…**
6. Input the appropriate hypothesized value in the box next to **Hypothesized difference:** (this is usually 0)
7. Select the appropriate alternative from the dropdown menu next to **Alternative hypothesis:**
8. Select **Use the pooled estimate of the proportion** from the drop down menu next to **Test method:**
9. Click **OK**
10. Click **OK**

Raw data

1. Input the raw data two separate columns
2. Click **Stat** then click **Basic Statistics** then click **2 Proportion…**
3. Select **Each sample is in its own column** from the dropdown menu
4. Click in the box next to **Sample 1:**
5. Double-click the column name where the first group's raw data is stored
6. Click in the box next to **Sample 2:**
7. Double-click the column name where the second group's raw data are stored
8. Click **Options…**
9. Input the appropriate hypothesized value in the box next to **Hypothesized difference:** (this is usually 0)
10. Select the appropriate alternative from the dropdown menu next to **Alternative hypothesis:**
11. Select **Use the pooled estimate of the proportion** from the drop down menu next to **Test method:**
12. Click **OK**
13. Click **OK**

SPSS

SPSS does not have the functionality to automatically produce a *z* test for the difference of two proportions.

Excel

Excel does not have the functionality to automatically produce a *z*-test for the difference of two proportions. However, you can type the formulas into a cell for the test statistic in order to have Excel calculate this for you. Then use the methods from Chapter 6 to find the *P*-value using the Normal distribution.

TI-83/84

1. Press the **STAT** key
2. Highlight **TESTS**
3. Highlight **2-PropZTest…** and press **ENTER**
4. Next to **x1** type the number of successes from the first sample
5. Next to **n1** type the sample size from the first sample
6. Next to **x2** type the number of successes from the second sample
7. Next to **n2** type the sample size from the second sample
8. Next to **p1,** highlight the appropriate alternative hypothesis
9. Highlight **Calculate** and press **ENTER**

TI-Nspire

1. Under **Scratchpad** select **A Calculate**
2. Press the **menu** key then select **6:Statistics** then select **7:Stat Tests** then **6:2-Prop z Test…** then press **enter**
3. In the box next to **Successes, x1** type the number of successes from the first sample
4. In the box next to **n1** type the number of trials from the first sample
5. In the box next to **Successes, x2** type the number of successes from the second sample
6. In the box next to **n2** type the sample size from the second sample
7. In the box next to **Alternate Hyp** choose the appropriate alternative hypothesis from the drop-down menu
8. Press **OK**

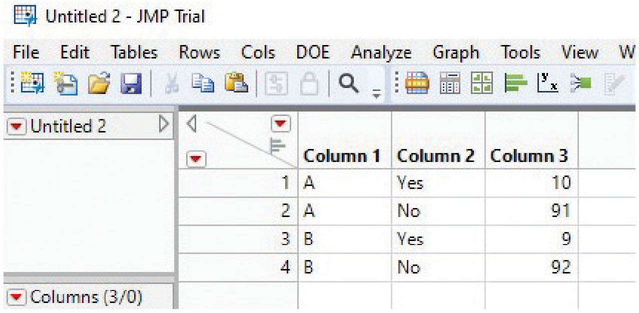
Confidence Interval for $p_1 - p_2$

JMP

Summarized data

1. Input the data into the JMP data table with categories for one variable in the first column, categories for the second variable in the second column, and counts for each combination in the third column

Copyright 2025 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s). Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.



	Column 1	Column 2	Column 3
1	A	Yes	10
2	A	No	91
3	B	Yes	9
4	B	No	92

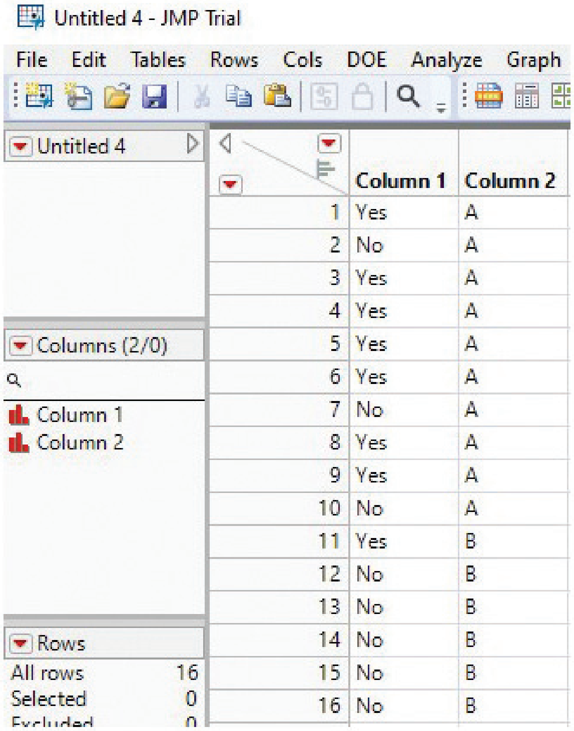
Source: JMP

2. Click **Analyze** and select **Fit Y by X**
3. Click and drag the first column containing the response variable from the box under **Select Columns** to the box next to **X, Factor**
4. Click and drag the second column containing the group information from the box under **Select Columns** to the box next to **Y, Response**
5. Click and drag the third column containing the counts for each combination from the box under **Select Columns** to the box next to **Freq**
6. Click **OK**
7. Click the red arrow next to **Contingency Analysis of…** and select **Two Sample Test for Proportions**

Note: You can change the response of interest (i.e., “Yes” instead of “No” or “Success” instead of “Failure”) by clicking the radio button at the bottom of the **Two Sample Test for Proportions** section.

Raw data

1. Input the raw data into two separate columns: one containing the response variable and one containing the group information
2. Click **Analyze** and select **Fit Y by X**



	Column 1	Column 2
1	Yes	A
2	No	A
3	Yes	A
4	Yes	A
5	Yes	A
6	Yes	A
7	No	A
8	Yes	A
9	Yes	A
10	No	A
11	Yes	B
12	No	B
13	No	B
14	No	B
15	No	B
16	No	B

Source: Microsoft Corporation

3. Click and drag the first column containing the response variable from the box under **Select Columns** to the box next to **Y, Response**
4. Click and drag the second column containing the group information from the box under **Select Columns** to the box next to **X, Factor**
5. Click **OK**
6. Click the red arrow next to **Contingency Analysis of…** and select **Two Sample Test for Proportions**

Note: You can change the response of interest (i.e., “Yes” instead of “No” or “Success” instead of “Failure”) by clicking the radio button at the bottom of the **Two Sample Test for Proportions** section.

Minitab

Summarized data

1. Click **Stat** then click **Basic Statistics** then click **2 Proportions…**
2. Select **Summarized data** from the dropdown menu
3. Summary statistics for the first sample will be entered in the **Sample 1** column. Type the number of successes in the **Number of events:** row. Type the sample size in the **Number of trials:** row.
4. Summary statistics for the second sample will be entered in the **Sample 2** column. Type the number of successes in the **Number of events:** row. Type the sample size in the **Number of trials:** row.
5. Click **Options…**
6. Input the appropriate confidence level in the box next to **Confidence level:**
7. Click **OK**
8. Click **OK**

Raw data

1. Input the raw data two separate columns
2. Click **Stat** then click **Basic Statistics** then click **2 Proportion…**
3. Select **Each sample is in its own column** from the dropdown menu
4. Click in the box next to **Sample 1:**
5. Double click the column name where the first group’s raw data is stored
6. Click in the box next to **Sample 2:**
7. Double click the column name where the second group’s raw data is stored
8. Click **Options…**
9. Input the appropriate confidence level in the box next to **Confidence level:**
10. Click **OK**
11. Click **OK**

SPSS

SPSS does not have the functionality to automatically produce a confidence interval for the difference of two proportions.

Excel

Excel does not have the functionality to automatically produce a confidence interval for the difference of two proportions. However, you can type the formulas into two separate cells for the lower and upper limit to have Excel calculate these results for you.

TI-83/84

1. Press the **STAT** key
2. Highlight **TESTS**
3. Highlight **2-PropZInt…** and press **ENTER**
4. Next to **x1** type the number of successes from the first sample
5. Next to **n1** type the sample size from the first sample
6. Next to **x2** type the number of successes from the second sample
7. Next to **n2** type the sample size from the second sample
8. Next to **C-Level** type the appropriate confidence level
9. Highlight **Calculate** and press **ENTER**

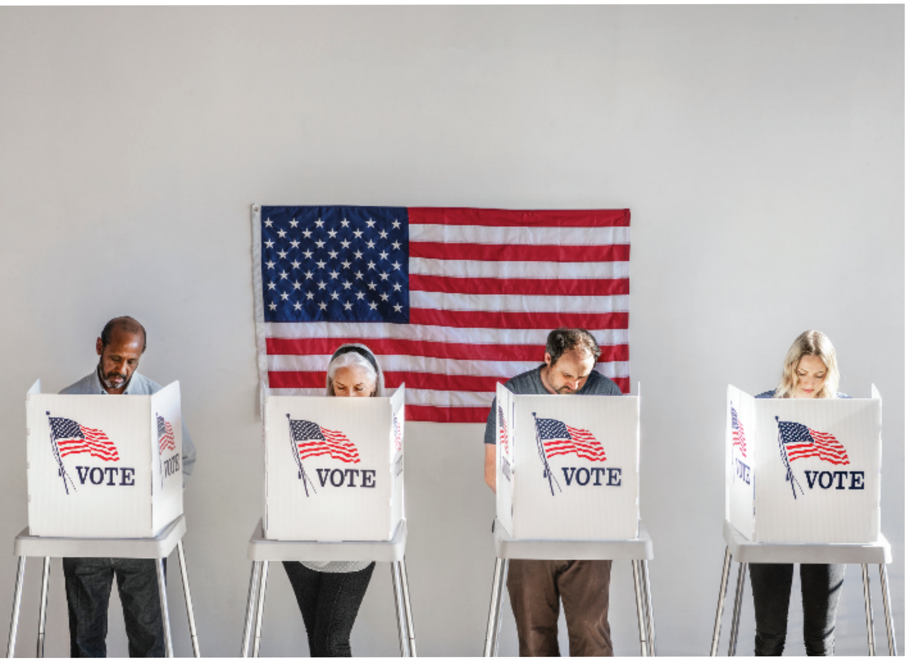
TI-Nspire

1. Under **Scratchpad** select **A Calculate**
2. Press the **menu** key then select **6:Statistics** then select **6:Confidence Intervals** then **6:2-Prop z Interval…** then press enter
3. In the box next to **Successes, x1** type the number of successes from the first sample
4. In the box next to **n1** type the number of trials from the first sample
5. In the box next to **Successes, x2** type the number of successes from the second sample
6. In the box next to **n2** type the number of trials from the second sample
7. In the box next to **C Level** input the appropriate confidence level
8. Press **OK**

Chapter

12

The Analysis of Categorical Data and Goodness-of-Fit Tests



Rawpixel.com/Shutterstock.com

It is often the case that information is collected on categorical variables, such as political affiliation, gender, or college major. As with numerical data, categorical data sets can be univariate (consisting of observations on a single categorical variable), bivariate (observations on two categorical variables), or even multivariate. In this chapter, methods for analyzing univariate categorical data sets are considered in Section 11.1 and methods appropriate for use with bivariate categorical data are introduced in Section 11.2.

Learning Objectives

After completing this chapter, you will be able to:

- Explain the differences between goodness-of-fit tests, tests for homogeneity, and tests of independence.
- Test hypotheses about the distribution of a categorical variable using a goodness-of-fit test and interpret the results in context.
- Test the hypothesis that the distribution of a categorical variable is the same for two or more populations or treatments using a test of homogeneity and interpret the results in context.
- Test hypotheses about association between two categorical variables using a test of independence and interpret the results in context.

Section 12.1 Chi-Square Tests for Univariate Data

Univariate categorical data sets arise in a variety of settings. If each student in a sample of 100 is classified according to whether they are enrolled full-time or part-time, data on a categorical variable with two categories result. Each registered voter in a sample of 100 selected from those registered in a particular city might be asked which of the five city council members they favor for mayor. This would yield observations on a categorical variable with five categories.

Univariate categorical data are conveniently summarized in a **one-way frequency table**. For example, many colleges now allow students to pay their tuition using a credit card (although paying by credit card may not be the best option for students–for a discussion on the pros and cons, see **"Paying College Tuition with a Credit Card," edvisors.com/plan-for-college/paying-for-college/pay-tuition-with-a-credit-card/, May 23, 2022**). Suppose that 100 randomly selected students at one of these colleges participated in a survey, with possible responses being definitely will use a credit card to pay tuition next year, probably will use a credit card, probably won't use a credit card, and definitely won't use a credit card. The first few observations might be

Probably will   Definitely will not   Probably will not
Probably will not   Definitely will   Definitely will not

Counting the number of observations of each type might then result in the following one-way table:

	Response			
	Definitely Will	**Probably Will**	**Probably Will Not**	**Definitely Will Not**
Frequency	14	12	24	50

For a categorical variable with k possible values (k different categories), sample data are summarized in a one-way frequency table consisting of k cells, which can be displayed either horizontally or vertically. The term cell refers to an individual entry in the table, so the table above would be described as having four cells.

In this section, you will see how to test hypotheses about the proportions of the population that fall into the possible categories. For example, a college that allows students to pay tuition by credit card might be interested in determining whether the four possible responses to the credit card question occur equally often in the population of all students at the college. If this is indeed the case, the long-run proportion of responses falling into each of the four categories is 1/4, or 0.25. The test procedure to be presented shortly would allow the college to decide whether it is plausible that the population proportions for all four categories are equal to 0.25.

Notation and Hypotheses

k = number of categories of a categorical variable

p_1 = population proportion for Category 1

p_2 = population proportion for Category 2

$\vdots$

p_k = population proportion for Category k

(Note: $p_1 + p_2 + \cdots + p_k = 1$)

The hypotheses to be tested have the form

H_0: p_1 = hypothesized proportion for Category 1
p_2 = hypothesized proportion for Category 2
$\vdots$
p_k = hypothesized proportion for Category k

H_a: H_0 is not true, so at least one of the population category proportions differs from the corresponding hypothesized value.

Copyright 2025 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s). Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

For the example using responses to the tuition survey, you could use

p_1 = the proportion of all students who will definitely pay by credit card
p_2 = the proportion of all students who will probably pay by credit card
p_3 = the proportion of all students who will probably not pay by credit card

and

p_4 = the proportion of all students who will definitely not pay by credit card

The null hypothesis of interest is then

$$H_0\colon\ p_1 = 0.25, p_2 = 0.25, p_3 = 0.25, p_4 = 0.25$$

A null hypothesis of the type just described can be tested by first selecting a random sample of size n and then classifying each sample response into one of the k possible categories. To decide whether the sample data are compatible with the null hypothesis, you compare the observed cell counts (frequencies) to the cell counts that would have been expected when the null hypothesis is true. The expected cell counts are

Expected cell count for Category 1 = np_1
Expected cell count for Category 2 = np_2

and so on. The expected cell counts when H_0 is true result from using the corresponding hypothesized proportions to calculate the expected counts.

Example 12.1 Jeopardy! Viewers

Amanda Edwards/Sringer/Getty Images Entertainment /Getty Images

The article **"Memo to Alex Trebek: Your Viewers Are Nerds" (yougov.com, October 21, 2016)** reports that viewers of the popular game show *Jeopardy!* tend to label themselves as nerds and that *Jeopardy!* viewers are more educated than the average American. These conclusions were based on a survey of a representative sample of people who said that they had watched the show in the past year. Data on the highest level of education completed for the sample of *Jeopardy!* viewers that are consistent with percentages given in the article are given in the accompanying table. The article did not give the sample size, so the numbers in the table are based on assuming a sample size of 1000.

Highest Level of Education Completed	Observed Frequency
Less than high school	10
High school	270
Some college	180
Two-year degree	70
Four-year college degree	300
Post-graduate study	170
Total	**1000**

The article also indicated that for the general population in the United States, the percentage of the population falling into each of the education categories is as shown in the following table.

Highest Level of Education Completed	Percentage of U.S. Population
Less than high school	7%
High school	36%
Some college	22%
Two-year degree	9%
Four-year college degree	17%
Post-graduate study	9%
Total	**100%**

The given information can be used to investigate whether the claim that *Jeopardy!* viewers tend to be more educated than the general population. You can represent the education category proportions as

p_1 = proportion of *Jeopardy!* viewers with a highest level of education completed of less than high school
p_2 = proportion of *Jeopardy!* viewers with a highest level of education completed of high school
p_3 = proportion of *Jeopardy!* viewers with a highest level of education completed of some college
p_4 = proportion of *Jeopardy!* viewers with a highest level of education completed of two-year degree
p_5 = proportion of *Jeopardy!* viewers with a highest level of education completed of four-year degree
p_6 = proportion of *Jeopardy!* viewers with a highest level of education completed of post-graduate study

If *Jeopardy!* viewers are like the population in general in terms of education level, you would expect that the *Jeopardy!* viewer proportions would match those of the general population. This means that they would be

$P_1 = 0.07$
$P_2 = 0.36$
$P_3 = 0.22$
$P_4 = 0.09$
$P_5 = 0.17$
$P_6 = 0.09$

What Do You Think?

What null hypothesis would you test if you wanted to know if *Jeopardy*! viewers are like the general population in terms of education level?

The hypotheses of interest are then

$$H_0: p_1 = 0.07, \quad p_2 = 0.36, \quad p_3 = 0.22, \quad p_4 = 0.09, \quad p_5 = 0.17, \quad p_6 = 0.09$$
$$H_a: H_0 \text{ is not true.}$$

There were a total of 1000 *Jeopardy!* viewers in the sample. If the null hypothesis is true, the expected counts for the first two categories are

$$\begin{pmatrix}\text{expected count for}\\ \text{less than high school}\end{pmatrix} = n\begin{pmatrix}\text{hypothesized proportion}\\ \text{for less than high school}\end{pmatrix} = 1000(0.07) = 70$$

$$\begin{pmatrix}\text{expected count for}\\ \text{high school}\end{pmatrix} = n\begin{pmatrix}\text{hypothesized proportion}\\ \text{for high school}\end{pmatrix} = 1000(0.36) = 360$$

Expected counts for the other four categories are calculated in a similar way. The observed and expected counts are given in the following table.

Highest Level of Education Completed	Observed Count	Expected Count
Less than high school	10	70
High school	270	360
Some college	180	220
Two-year degree	70	90
Four-year college degree	300	170
Post-graduate study	170	90
Total	**1000**	**1000**

Because the observed counts are based on a *sample* of *Jeopardy!* viewers, it would be surprising to see *exactly* 7% of the sample falling into the first education category, exactly 36% falling into the second category, and so on, even when H_0 is true. If the differences between the observed and expected cell counts can reasonably be attributed to sampling variability, the data are considered compatible with H_0. On the other hand, if the

Copyright 2025 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s). Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

differences between the observed and the expected cell counts are too large to be explained by chance differences from one sample to another, H_0 should be rejected in favor of H_a. To make a decision, you need to assess how different the observed and expected counts are.

The goodness-of-fit statistic, denoted by X^2, is a measure of the extent to which the observed counts differ from those expected when H_0 is true. (The Greek letter χ is often used in place of X. The symbol X^2 is referred to as the chi-square [χ^2] statistic. Using X^2 rather than χ^2 follows the convention of denoting sample statistics by Roman letters.)

For a sample of size n,

$$\begin{pmatrix}\text{expected cell} \\ \text{count}\end{pmatrix} = n\begin{pmatrix}\text{hypothesized value of corresponding} \\ \text{population proportion}\end{pmatrix}$$

The **goodness-of-fit statistic**, X^2, results from first calculating the quantity

$$\frac{(\text{observed cell count} - \text{expected cell count})^2}{\text{expected cell count}}$$

for each cell.

The X^2 statistic is the sum of these quantities for all k cells in the table:

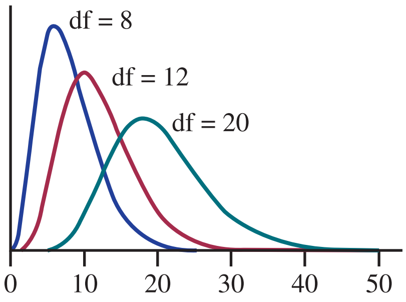
$$X^2 = \sum_{\text{all cells}} \frac{(\text{observed cell count} - \text{expected cell count})^2}{\text{expected cell count}}$$

The value of the X^2 statistic reflects the magnitude of the differences between observed and expected cell counts. When the differences are large, the value of X^2 tends to be large. This means that large values of X^2 suggest rejection of H_0. A small value of X^2 (it can never be negative) occurs when the observed cell counts are quite similar to those expected when H_0 is true, and would be consistent with H_0.

As with previous test procedures, a conclusion is reached by comparing a P-value to the significance level for the test. The P-value is calculated as the probability of observing a value of X^2 at least as large as the observed value when H_0 is true. This requires information about the sampling distribution of X^2 when H_0 is true.

When the null hypothesis is true and the sample size is large, the behavior of X^2 is described approximately by a **chi-square distribution**. A chi-square distribution curve has no area associated with negative values. It is not symmetric and has a longer tail on the right. There are many chi-square distributions, and each one has a different number of degrees of freedom. Curves corresponding to several chi-square distributions are shown in Figure 12.1.

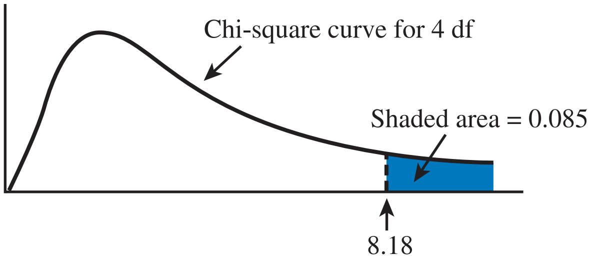
Figure 12.1
Chi-square curves.



For a test procedure based on the X^2 statistic, the associated P-value is the area under the appropriate chi-square curve and to the right of the calculated X^2 value. This area can be determined using technology or using Appendix Table 8, which gives upper-tail areas for chi-square distributions with up to 20 df.

To find the area to the right of a particular X^2 value using Appendix Table 8, locate the appropriate df column in the table. Determine which listed value is closest to the X^2 value of interest, and read the right-tail area corresponding to this value from the left-hand column of the table. For example, for a chi-square distribution with df = 4, the area to the right of $X^2 = 8.18$ is 0.085, as shown in Figure 12.2. For this same chi-square distribution (df = 4), the area to the right of 9.70 is approximately 0.045 (the area to the right of 9.74, the closest entry in the table for df = 4).

Figure 12.2
A chi-square upper-tail area.



It is also possible to use a statistical software package or a graphing calculator to determine areas under a chi-square distribution curve.

Goodness-of-Fit Tests

When H_0 is true, the X^2 goodness-of-fit statistic has approximately a chi-square distribution with df = $(k - 1)$, as long as none of the expected cell counts are too small. When expected counts are small, and especially when an expected count is less than 1, the value of $\frac{(\text{observed cell count} - \text{expected cell count})^2}{\text{expected cell count}}$ can be inflated because it involves dividing by a small number. The use of the chi-square distribution is appropriate when the sample size is large enough for every expected cell count to be at least 5. If any of the expected cell frequencies are less than 5, categories can be combined in a sensible way to create acceptable expected cell counts. If you do this, remember to calculate the number of degrees of freedom based on the reduced number of categories.

X^2 Goodness-of-Fit Test

Hypotheses: H_0: p_1 = hypothesized proportion for Category 1

$\vdots$

p_k = hypothesized proportion for Category k

H_a: H_0 is not true. At least one of the education category proportions is different from the value specified in the null hypothesis.

Test Statistic:

$$X^2 = \sum_{\text{all cells}} \frac{(\text{observed cell count} - \text{expected cell count})^2}{\text{expected cell count}}$$

***P*-values:** When H_0 is true and all expected counts are at least 5, X^2 has approximately a chi-square distribution with df = $k - 1$. The *P*-value associated with the calculated test statistic value is the area to the right of X^2 under the df = $k - 1$ chi-square curve. Upper-tail areas for chi-square distributions are found in Appendix Table 8 or can be found using technology.

Assumptions:
1. Observed cell counts are based on a random sample.
2. The sample size is large. The sample size is large enough for the chi-square test to be appropriate as long as every expected cell count is at least 5.

Copyright 2025 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s). Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

Example 12.2 *Jeopardy!* Viewers Revisited

● You can use the *Jeopardy!* data in Example 12.1 to test the null hypothesis that *Jeopardy!* viewers are like the general population in terms of education level. A significance level of 0.05 will be used. The observed and expected counts calculated in Example 12.1 are given in the accompanying table.

Highest Level of Education Completed	Observed Count	Expected Count
Less than high school	10	70
High school	270	360
Some college	180	220
Two-year degree	70	90
Four-year college degree	300	170
Post-graduate study	170	90
Total	**1000**	**1000**

Understand the context ➤

1. Let p_1, p_2, p_3, p_4, p_5, and p_6 denote the proportions falling into the six education categories as defined in Example 12.1.

If *Jeopardy!* viewers are like the population in terms of level of education, then you would expect the proportions falling into the education categories to match those for the general population. This leads to the null hypothesis given in Example 12.1:

2. H_0: $p_1 = 0.07, p_2 = 0.36, p_3 = 0.22, p_4 = 0.09, p_5 = 0.17, p_6 = 0.09$
3. H_a: H_0 is not true. At least one of the education category proportions is different from the value specified in the null hypothesis.
4. Significance level: $\alpha = 0.05$.

Formulate a plan ➤

5. Test statistic: $$X^2 = \sum_{\text{all cells}} \frac{(\text{observed cell count} - \text{expected cell count})^2}{\text{expected cell count}}$$
6. Assumptions: The expected cell counts (from Example 12.1) are all greater than 5. The viewers were a representative sample of *Jeopardy!* viewers.

Do the work ➤

7. Calculation:

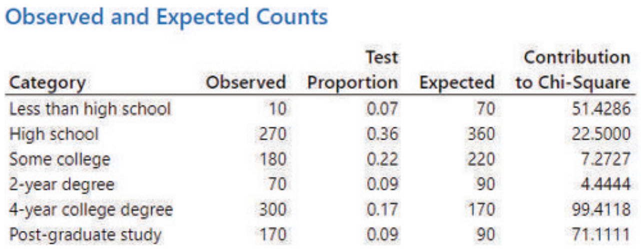
$$X^2 = \frac{(10-70)^2}{70} + \frac{(270-360)^2}{360} + \frac{(180-220)^2}{220} + \frac{(70-90)^2}{90} + \frac{(300-170)^2}{170}$$
$$+ \frac{(170-90)^2}{90}$$
$$= 51.429 + 22.500 + 7.273 + 4.444 + 99.412 + 71.111$$
$$= 256.169$$

8. P-value: The P-value is based on a chi-square distribution with df $= 6 - 1 = 5$. The calculated value of X^2 is greater than 20.51 (the largest entry in the df $= 5$ column of Appendix Table 8), so P-value < 0.001.

Interpret the result ➤

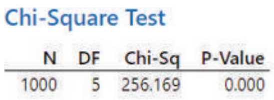
9. Conclusion: Because the P-value is less than α, H_0 is rejected. There is convincing evidence that *Jeopardy!* viewers are not like the general population with respect to education level. Notice that the observed counts are much greater than the expected counts for the 4-year degree and post-graduate study categories. This is consistent with the conclusion in the article that *Jeopardy!* viewers are "more educated than the average American."

Statistical software can be used to perform a chi-square goodness-of-fit test. Minitab output for the data and hypothesized proportions of this example is shown here.



```
Observed and Expected Counts

                                  Test                Contribution
Category                Observed  Proportion  Expected  to Chi-Square
Less than high school         10        0.07        70        51.4286
High school                  270        0.36       360        22.5000
Some college                 180        0.22       220         7.2727
2-year degree                 70        0.09        90         4.4444
4-year college degree        300        0.17       170        99.4118
Post-graduate study          170        0.09        90        71.1111
```

● **Data set available online**



```
Chi-Square Test

   N  DF   Chi-Sq  P-Value
1000   5  256.169    0.000
```

Example 12.3 Tasty Dog Food?

● The article **"Can People Distinguish Pâté from Dog Food?" (American Association of Wine Economists, April 2009, wine-economics.org)** describes a study that investigated whether people can tell the difference between dog food, pâté (a spread made of finely chopped liver, meat, or fish), and processed meats (such as Spam and liverwurst).

Researchers used a food processor to make spreads that had the same texture and consistency as pâté from Newman's Own dog food and from the processed meats. Each participant in the study tasted five spreads (duck liver pâté, Spam, dog food, pork liver pâté, and liverwurst). After tasting all five spreads, each participant was asked to choose the one that they thought was the dog food. The researchers recorded which of the five spreads was selected as the one the participant thought was dog food. Consider the data in the following table.

	Spread Chosen as Dog Food				
	Duck liver pâté	**Spam**	**Dog food**	**Pork liver pâté**	**Liverwurst**
Frequency	3	11	8	6	22

(Note: The frequencies in the table are consistent with summary values given in the paper. However, the sample size in the study was not actually 50.)

You can use the dog food taste data to test the hypothesis that the five different spreads (duck liver pâté, Spam, dog food, pork liver pâté, and liverwurst) are chosen equally often when people who have tasted all five spreads are asked to identify the one that they think is the dog food. If this is the case, each category proportion would be $1/5 = 0.20$. Because the sample size was 50, the expected count for the duck liver category is $50(0.20) = 10$. Because the other hypothesized category proportions are also 0.20, all of the expected counts are equal to 10.

Expected counts:

Category 1: $np_1 = 50(0.20) = 10$
Category 2: $np_2 = 50(0.20) = 10$
Category 3: $np_3 = 50(0.20) = 10$
Category 4: $np_4 = 50(0.20) = 10$
Category 5: $np_5 = 50(0.20) = 10$

You are now ready to use the nine-step hypothesis testing procedure to test the hypotheses of interest.

Understand the context ➤

1. p_1 = proportion of all people who would choose duck liver pâté as the dog food
   p_2 = proportion of all people who would choose Spam as the dog food
   p_3 = proportion of all people who would choose dog food as the dog food
   p_4 = proportion of all people who would choose pork liver pâté as the dog food
   p_5 = proportion of all people who would choose liverwurst as the dog food
2. $H_0: p_1 = 0.20, p_2 = 0.20, p_3 = 0.20, p_4 = 0.20, p_5 = 0.20$
3. H_a: At least one of the population proportions is not 0.20.
4. Significance level: $\alpha = 0.05$.

Formulate a plan ➤

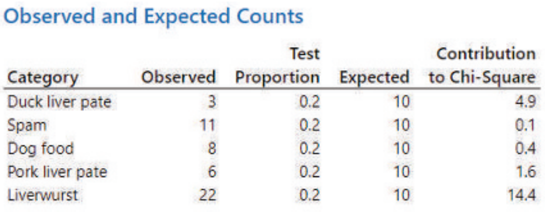
5. Test statistic: $$X^2 = \sum_{\text{all categories}} \frac{(\text{observed count} - \text{expected count})^2}{\text{expected count}}$$
6. Assumptions: In order to use the chi-square goodness-of-fit test, you must be willing to assume that the participants in this study can be regarded as a random or

● **Data set available online**

Copyright 2025 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s). Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

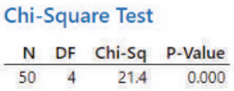
representative sample. If this assumption is not reasonable, you should be very careful generalizing results from this analysis to any larger population. All expected counts are at least 5, so the sample size is large enough for the chi-square goodness-of-fit test to be appropriate.

Do the work ➤ **7.** Calculations: From Minitab



Observed and Expected Counts

Category	Observed	Test Proportion	Expected	Contribution to Chi-Square
Duck liver pate	3	0.2	10	4.9
Spam	11	0.2	10	0.1
Dog food	8	0.2	10	0.4
Pork liver pate	6	0.2	10	1.6
Liverwurst	22	0.2	10	14.4



Chi-Square Test

N	DF	Chi-Sq	P-Value
50	4	21.4	0.000

Interpret the results ➤ **8.** *P*-value: From Minitab, the *P*-value is 0.000.

9. Conclusion: Based on the sample data, there is convincing evidence that the proportion identifying a spread as dog food is not the same for all five spreads. Here, it is interesting to note that the large differences between the observed counts and the counts that would have been expected if the null hypothesis of equal proportions was true are in the duck liver pâté and the liverwurst categories, indicating that fewer than expected chose the duck liver and many more than expected chose the liverwurst as the one they thought was dog food. So, although you reject the hypothesis that proportion choosing each of the five spreads is the same, it is not because people were actually able to identify which one was really dog food!

Exercises 12.1–12.13

● **Data set available online**

12.1 A particular cell phone case is available in a choice of four different colors. A store sells all four colors. To test the hypothesis that sales are equally divided among the four colors, a random sample of 100 purchases is identified.

a. If the resulting X^2 value were 6.4, what is the conclusion when using a test with significance level 0.05?

b. What conclusion would be appropriate at significance level 0.01 if $X^2 = 15.3$?

c. If there were six different covers rather than just four, what is the conclusion if $X^2 = 13.7$ and a test with $\alpha = 0.05$ was used?

12.2 From the given information in each case below, use Appendix Table 8 or technology to find the *P*-value for a chi-square test and give the conclusion for a significance level of $\alpha = 0.01$.

a. $X^2 = 7.5$, df $= 2$

b. $X^2 = 13.0$, df $= 6$

c. $X^2 = 18.0$, df $= 9$

d. $X^2 = 21.3$, df $= 4$

e. $X^2 = 5.0$, df $= 3$

12.3 ● In 2020, the University of Houston carried out a study for the Texas Lottery Commission **("Demographic Survey of Texas Lottery Players," 2020)**. The accompanying table gives the age distribution for a representative sample of 1631 Texas Lottery players.

Age Group	Frequency
18 to 24	198
25 to 34	238
35 to 44	243
45 to 54	265
55 to 64	250
65 and older	437
Total	**1631**

Using data from the U.S. Census Bureau (**www.census.gov**) for 2020, the age distribution of adults in Texas was: 11% between age 18 and 24, 20% between age 25 and 34, 19% between age 35 and 44, 17% between age 45 and 54, 16% between age 55 and 64, and 17% age 65 or older. Is it reasonable to conclude that one or more of these age groups buys a disproportionate share of Texas Lottery tickets? Use a chi-square goodness-of-fit test with $\alpha = 0.05$.

12.4 The **"Global Automotive 2020 Color Popularity Report" (Axalta Coating Systems, 2020)** included data on the colors for a sample of new cars sold in North America. The report stated that 38% of the cars in the sample were white, 19% were black, 15% were gray, 9% were silver, 7% were blue, 5% were red, and 7% were some other color. Suppose that these percentages were based on a random sample of 1200 new cars sold in North America. Is there convincing evidence that the proportions of new cars sold are not the same for all of the seven color categories?

12.5 ● The article **"1 in 3 Won't Consider a New Job Unless They Can Work Remotely" (StudyFinds, March 13, 2022)** summarizes data from a survey of 2000 people who work remotely or in a hybrid-remote mode (partially remote). Responses to a question about preferred work mode are summarized in the accompanying table.

Entirely In-Person	Hybrid-Remote (Partially Remote)	Entirely Remote
740	780	480

a. Assume that this sample of 2000 workers is representative of the population of all workers who work remotely or in a hybrid-remote mode. Carry out a hypothesis test to determine if there is convincing evidence that these three work modes are not equally preferred.

b. Write a few sentences describing how the observed counts for the three work modes differ from what would be expected if the three work modes were equally preferred.

12.6 ● The authors of the paper **"External Factors and the Incidence of Severe Trauma: Time, Date, Season and Moon" (*Injury* [2014]: S93–S99)** classified admissions to hospitals in Germany according to season. They wondered if severe trauma injuries were more common in some seasons than others. Assume that there were 1200 trauma cases in the sample and that the sample is representative of severe trauma injuries in Germany. The data in the accompanying table are consistent with summary quantities given in the paper. Do these data support the theory that the proportion of severe trauma cases is not the same for the four seasons? Test the relevant hypotheses using a significance level of 0.05.

Season				
Winter	Spring	Summer	Fall	Total
228	332	352	288	1200

12.7 ● The **"2019 Young Money Survey" (Ameritrade, June 2019)** asked each person in a representative sample of 1026 millennials (those born between 1981 and 1996) to respond to the following question: "Financially speaking, do you expect to be more successful, equally successful, or less successful than your parents?" Responses are summarized in the accompanying table. Is there convincing evidence that the response proportions are not all the same for the population of millennials? Use a significance level of 0.01.

More Successful	As Successful	Less Successful
453	359	133

12.8 ● The article **"2023 Student Money Survey" (wallethub.com, May 9, 2023)** summarizes the results of a "nationally representative online survey" of 200 college students. One question each student in the sample was asked was "What is your biggest post-graduation fear?" The responses are summarized in the accompanying table.

Biggest Fear	Observed Frequency
Student loan debt	67
Not finding a job	63
Credit card debt	36
Living with parents	34

a. Do these data provide convincing evidence that the four response proportions are not equal in the population of all college students? Test the relevant hypotheses using a significance level of $\alpha = 0.05$.

b. Write a few sentences commenting on what you learn from comparing the observed and expected frequencies calculated in Part (a).

12.9 The survey of college students described in the article referenced in the previous exercise also gave data on student responses to a question about what form of payment they use most often. The responses are summarized in the accompanying table. Suppose that someone has claimed that the category proportions in the population of all college students are 0.5 for credit cards, 0.3 for debit cards, and 0.1 for each of cash and the "other" response category. Do the sample data provide convincing evidence that this claim should be rejected? Test the relevant hypotheses using a significance level of $\alpha = 0.01$.

Form of Payment	Observed Frequency
Credit card	104
Debit card	66
Cash	24
Other	6

Copyright 2025 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s). Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

12.10 ● The article **"Many of the World's Billionaires Share a Zodiac Sign, Study Finds" (fortune.com, February 7, 2023)** gives information on the zodiac signs of 300 billionaires. The values in the accompanying table are consistent with percentages given in the article.

a. Is the claim that more billionaires are born under some zodiac signs than others justified? Carry out an appropriate hypothesis test using $\alpha = 0.05$.

b. Based on your conclusion in Part (a), what would be a better title for an article based on these data?

Zodiac Sign	Observed Frequency
Aries	24
Taurus	30
Gemini	24
Cancer	22
Leo	27
Virgo	24
Libra	36
Scorpio	18
Sagittarius	23
Capricorn	17
Aquarius	22
Pisces	33

12.11 How would you answer the following question:

Next Wednesday's meeting has been moved forward 2 days. What day is the meeting now that it has been rescheduled?

This question is ambiguous as "moved forward" can be interpreted in two different ways. Did you answer Monday or Friday? The authors of the paper **"Even Abstract Motion Influences the Understanding of Time" (*Metaphor and Symbol* [2011]: 260–271)** wondered if the answers Monday and Friday would be provided an equal proportion of the time. A sample of students at Stanford University were asked this question, and the responses are summarized in the following table.

Response	Frequency
Monday	11
Friday	33

The authors of the paper used a chi-squared goodness-of-fit test to test the null hypothesis H_0: $p_1 = 0.50$, $p_2 = 0.50$, where p_1 is the proportion who would respond Monday and p_2 is the proportion who would respond Friday. They reported $X^2 = 11.00$ and P-value < 0.001. What conclusion can be drawn from this test?

12.12 ● The article **"Current EV Registrations in the US: How Does Your State Stack Up and Who Grew the Most YOY?" (electrek.com, August 24, 2022)** reported the top five states for sales of hybrid cars in 2021 as California, Florida, Texas, Washington, and New York. Suppose that each car in a sample of 2021 hybrid car sales is classified by state where the sale took place. Sales from states other than the top five were excluded from the sample, resulting in the accompanying table.

State	Observed Frequency
California	230
Florida	55
Texas	45
Washington	40
New York	30
Total	**500**

(The given observed counts are artificial, but they are consistent with hybrid sales figures given in the article.)

The 2021 population estimates from the Census Bureau website are given in the accompanying table. The population proportion for each state was calculated by dividing each state population by the total population for all five states. Use the X^2 goodness-of-fit test and a significance level of $\alpha = 0.01$ to test the hypothesis that hybrid sales for these five states are proportional to the 2021 population for these states.

State	2021Population	Population Proportion
California	39,142,991	0.331
Florida	21,828,069	0.185
Texas	29,558,864	0.250
Washington	7,740,745	0.066
New York	19,857,492	0.168
Total	**118,128,161**	

12.13 A certain genetic characteristic of a particular plant can appear in one of three forms (phenotypes). A researcher has developed a theory, according to which the hypothesized proportions are $p_1 = 0.25$, $p_2 = 0.50$, and $p_3 = 0.25$. A random sample of 200 plants yields $X^2 = 4.63$.

a. Carry out a test of the null hypothesis that the theory is correct, using level of significance $\alpha = 0.05$.

b. Suppose that a random sample of 300 plants had resulted in the same value of X^2. How would the analysis and conclusion differ from those in Part (a)?

Section 12.2 Tests for Homogeneity and Independence in a Two-way Table

Bivariate categorial data (resulting from observations made on two different categorical variables) can also be summarized in a table. For example, suppose that a market researcher is interested in determining if there is any relationship between generation (Generation Z, millennials, Generation X, or Baby Boomers) and type of car purchased (gas engine, hybrid, or electric) for people who purchased a new car in the last year. The two variables of interest are *generation* and *car type*. If a random sample of 1000 people who purchased cars in the last year is selected and these two variables are recorded for each person in the sample, the data set is bivariate and might initially be displayed as follows:

Observation	Generation	Car Type
1	Gen X	Electric
2	Boomer	Electric
3	Gen X	Gas
⋮		
999	Millennial	Hybrid
1000	Gen Z	Gas

Bivariate categorical data of this sort can most easily be summarized by constructing a **two-way frequency table**, or **contingency table**. This is a rectangular table that consists of a row for each possible category of one variable (x) and a column for each possible category of the other variable (y). There is a cell in the table for each possible (x, y) category combination.

The number of times each particular (x, y) combination occurs in the data set is determined, and these numbers (frequencies) are entered in the corresponding cells of the table. The resulting numbers are called **observed cell counts**. The table for the example relating generation to car type contains 4 rows and 3 columns (because x has 4 possible values and y has 3 possible values). Table 12.1 is one possible table.

Table 12.1 An Example of a 4 × 4 Frequency Table

	Gas	Hybrid	Electric	Row Marginal Total
Gen Z	115	50	35	**200**
Millennials	155	65	40	**260**
Gen X	155	55	30	**240**
Boomers	185	100	15	**300**
Column Marginal Total	**610**	**270**	**120**	**1000**

Marginal totals are obtained by adding the observed cell counts in each row and also in each column of the table. The row and column marginal totals, along with the total of all observed cell counts in the table, the **grand total**, have been included in Table 12.1.

The marginal totals provide information about the distribution of observed values for each variable separately. In this example, the row marginal totals reveal that the sample consisted of 200 Generation *Z*, 260 millennials, 240 Generation *X*, and 300 Baby Boomers. Similarly, column marginal totals indicate how often each of the different car types were purchased: 610 were gas engines cars, 270 were hybrids, and 120 were electric. The grand total, 1000, is the total number of observations in the bivariate data set. This two-way table could be used to investigate any association between generation and type of car purchased.

Two-way frequency tables are often characterized by the number of rows and columns in the table (specified in that order: rows first, then columns). Table 12.1 is called a 4 × 3 table.

The smallest two-way frequency table is a 2 × 2 table, which has only two rows and two columns, resulting in four cells.

Two-way tables arise naturally in two different types of investigations. A researcher may be interested in comparing two or more populations or treatments on the basis of a single categorical variable and so might obtain independent samples from each population or treatment. For example, data could be collected at a university to compare students, faculty, and staff on the basis of primary mode of transportation to campus (car, bicycle, motorcycle, bus, or by foot). One random sample of 200 students, another random sample of 100 faculty members, and a third random sample of 150 staff members might be selected, and the individuals in each sample could be interviewed to obtain the necessary transportation information.

Data from such a study could be summarized in a 3 × 5 two-way frequency table with row categories of student, faculty, and staff and column categories corresponding to the five possible modes of transportation. The observed cell counts could then be used to learn about differences and similarities among the three groups with respect to mode of transportation. This type of bivariate categorical data set is characterized by having one set of marginal totals predetermined (the sample sizes for the different groups). In the 3 × 5 situation just discussed, the row totals would be fixed at 200, 100, and 150.

A two-way table also arises when the values of two different categorical variables are observed for all individuals or items in a single sample. For example, a sample of 500 registered voters might be selected. Each voter could then be asked both if they favored a particular property tax initiative and if they were a registered Democrat, Republican, or Independent. This would result in a bivariate data set with x representing the variable *political affiliation* (with categories Democrat, Republican, and Independent) and y representing the variable *response* (favors initiative or opposes initiative). The corresponding 3 × 2 frequency table could then be used to investigate any association between position on the tax initiative and political affiliation. This type of bivariate categorical data set is characterized by having only the grand total predetermined (by the sample size).

Comparing Two or More Populations or Treatments: A Test of Homogeneity

When the value of a categorical variable is recorded for members of independent random samples obtained from each of the populations or treatments under study, the question of interest is whether the category proportions are the same for all the populations or treatments. As in Section 12.1, the test procedure uses a chi-square statistic that compares the observed counts to those that would be expected if there were no differences.

Example 12.4 Heart Attacks in High-Rise Buildings Revisited

Understand the context ➤

● The paper **"Out-of-Hospital Cardiac Arrest in High-Rise Buildings: Delays to Patient Care and Effect on Survival" (*Canadian Medical Association Journal* [2016]: 413–419)** compared heart attack survival rates for people who live in a house or townhouse, people who live on the first or second floor of an apartment building, and people who live on the third or a higher floor in an apartment building. Table 12.2, a 3 × 4 two-way frequency table, is the result of classifying each heart attack victim in independently selected representative samples of 5531 heart attacks that occurred in a house or townhouse, 667 heart attacks that occurred in an apartment building on the first or second floor, and 1696 heart attacks that occurred in an apartment building on the third or higher floor into one of two survival categories (survived and did not survive).

Table 12.2 Observed Counts for Example 12.4

	Survived	Did Not Survive	Row Marginal Total
House or Townhouse	217	5314	**5531**
Apartment First or Second Floor	35	632	**667**
Apartment Third or Higher Floor	46	1650	**1696**
Column Marginal Total	**298**	**7596**	**7894**

● **Data set available online**

Notice that there are a total of 7894 heart attack victims in the three samples combined. Of these, 298 survived. The proportion of the total who survived is then

$$\frac{298}{7894} = 0.038$$

If there is no difference in survival for the three different groups, you would expect about 3.8% of the heart attack victims who live in a house or townhouse to survive, about 3.8% of the heart attack victims who live in an apartment on the first or second floor to survive, and about 3.8% of the heart attack victims who live in an apartment on the third or a higher floor to survive. This means that if there is no difference in survival, the expected number surviving for each of the three cells in the "survived" column are

Expected count for house or townhouse and survive = 0.038(5531) = 210.178
Expected count for apartment on first or second floor and survive = 0.038(667) = 23.346
Expected count for apartment on third or higher floor and survive = 0.038(1696) = 64.448

Notice that the expected cell counts do not need to be whole numbers.

The expected cell counts for the remaining cells can be calculated in a similar manner. The proportion of the total who did not survive is

$$\frac{7596}{7894} = 0.962$$

Then, if there is no difference in survival for the three groups,

Expected count for house or townhouse and did not survive = 0.962(5531) = 5320.822
Expected count for apartment on first or second floor and did not survive = 0.962(667) = 641.654
Expected count for apartment on third or higher floor and did not survive = 0.962(1696) = 1631.552

It is common practice to display the observed cell counts and the corresponding expected cell counts in the same table, with the expected cell counts enclosed in parentheses. Table 12.3 gives the observed cell counts and the expected cell counts. Notice that each marginal total for the expected cell counts is equal to the corresponding marginal total for the observed cell counts. (Sometimes there may be small differences in the marginal totals due to rounding when calculating the expected cell counts.)

Table 12.3 **Observed and Expected Counts for Example 12.4**

	Survived	Did Not Survive	Row Marginal Total
House or Townhouse	217 (210.178)	5314 (5320.822)	**5531**
Apartment First or Second Floor	35 (23.346)	632 (641.654)	**667**
Apartment Third or Higher Floor	46 (64.448)	1650 (1631.552)	**1696**
Column Marginal Total	**298**	**7596**	**7894**

A quick comparison of the observed and expected cell counts in Table 12.3 reveals some large differences, suggesting that the survival proportions may not be the same for the three groups considered. This will be explored further in Example 12.5.

In Example 12.4, the expected count for a cell corresponding to a particular group–survival combination was calculated in two steps. First, the survival *marginal proportion* was calculated (for example, 298/7894 for the "survived" category). Then this proportion was multiplied by a marginal group total (for example, 5531(298/7894) for the house or

Copyright 2025 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s). Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

townhouse group). Algebraically, this is equivalent to first multiplying the row and column marginal totals and then dividing by the grand total:

$$\frac{(5531)(298)}{7894}$$

To compare two or more populations or treatments on the basis of a categorical variable, calculate an **expected cell count** for each cell by selecting the corresponding row and column marginal totals and then calculating

$$\text{expected cell count} = \frac{(\text{row marginal total})(\text{column marginal total})}{\text{grand total}}$$

These quantities represent what would be expected when there is no difference between the groups under study.

The X^2 statistic, introduced in Section 12.1, can now be used to compare the observed cell counts to the expected cell counts. A large value of X^2 results when there are large differences between the observed and expected counts and suggests that the hypothesis of no differences between the populations should be rejected. A formal test procedure is described in the following box.

X^2 Test for Homogeneity

Null Hypothesis: H_0: The category proportions are the same for all the populations or treatments (homogeneity of populations or treatments).

Alternative Hypothesis: H_a: The category proportions are not all the same for all of the populations or treatments.

Test Statistic: $$X^2 = \sum_{\text{all cells}} \frac{(\text{observed cell count} - \text{expected cell count})^2}{\text{expected cell count}}$$

The expected cell counts are estimated from the sample data (assuming that H_0 is true) using the formula

$$\text{expected cell count} = \frac{(\text{row marginal total})(\text{column marginal total})}{\text{grand total}}$$

***P*-values:** When H_0 is true and the assumptions of the X^2 test for homogeneity are satisfied, X^2 has approximately a chi-square distribution with

$$\text{df} = (\text{number of rows} - 1)(\text{number of columns} - 1)$$

The *P*-value associated with the calculated test statistic value is the area to the right of X^2 under the chi-square curve with the appropriate df. Upper-tail areas for chi-square distributions are found using technology or Appendix Table 8.

Assumptions:

1. The data are from independently selected random samples or from subjects who were assigned at random to treatment groups.
2. The sample sizes are large: all expected counts are at least 5. If some expected counts are less than 5, rows or columns of the table may be combined to achieve a table with satisfactory expected counts.

Example 12.5 Heart Attacks in High-Rise Buildings Revisited

The following table of observed and expected cell counts appeared in Example 12.4.

	Survived	Did Not Survive	Row Marginal Total
House or Townhouse	217 (210.178)	5314 (5320.822)	**5531**
Apartment First or Second Floor	35 (23.346)	632 (641.654)	**667**
Apartment Third or Higher Floor	46 (64.448)	1650 (1631.552)	**1696**
Column Marginal Total	**298**	**7596**	**7894**

Suppose you are interested in learning if there is a difference in the survival category proportions for the three populations (heart attack victims who live in a house or townhouse, heart attack victims who live on the first or second floor of an apartment building, and heart attack victims who live on the third or a higher floor in an apartment building).

Understand the context ➤

Hypotheses: H_0: Proportions in each survival category are the same for all three groups.
H_a: The survival category proportions are not all the same for all three groups.

Significance level: A significance level of $\alpha = 0.05$ will be used.

Formulate a plan ➤

Test statistic: $$X^2 = \sum_{\text{all cells}} \frac{(\text{observed cell count} - \text{expected cell count})^2}{\text{expected cell count}}$$

Assumptions: The samples were independently selected and thought to be representative of the three populations of interest. This means that it is appropriate to use the chi-square test if the sample sizes are large enough. All of the expected cell counts are at least 5, so the samples are large enough to proceed with the test.

Do the work ➤

Calculation:

$$X^2 = \frac{(217 - 208.795)^2}{208.795} + \cdots + \frac{(1650 - 1631.976)^2}{1631.976} = 9.59$$

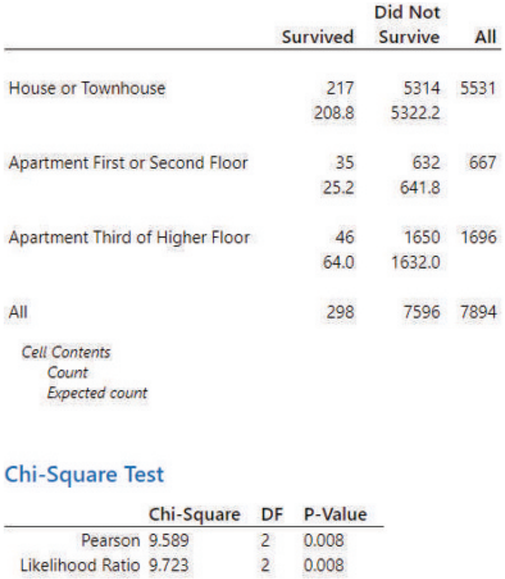
P-value: The two-way table for this example has 3 rows and 2 columns, so the appropriate number of df is $(3 - 1)(2 - 1) = 2$. The calculated value of the test statistic is between 9.21 and 10.59 in the 2-df column of Appendix Table 8, so $0.005 < P\text{-value} < 0.010$.

Interpret the results ➤

Conclusion: The *P*-value is less than α (0.05), so H_0 is rejected. There is convincing evidence that the proportions in the survival categories are not the same for the three groups compared. Notice that there are more people who survived in the house or townhouse and first or second floor apartment categories than would have been expected if the survival proportions were the same for all three groups. This led the researchers who collected these data to conclude that there is a smaller chance of survival for people who suffer a heart attack in an apartment that is on the third or higher floor.

Most statistical software packages can calculate expected cell counts, the value of the X^2 statistic, and the associated *P*-value. Minitab output for the data of Example 12.5 is shown here. The value of the test statistic and associated *P*-value for this test are the ones found in the row labeled "Pearson" in the Minitab output. Notice that Minitab provides an approximate *P*-value (0.008 for this example). Using Appendix Table 8, you could only obtain an interval that contained the *P*-value.

Copyright 2025 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s). Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.



```
                                              Did Not
                                   Survived   Survive   All

House or Townhouse                      217      5314  5531
                                      208.8    5322.2

Apartment First or Second Floor          35       632   667
                                       25.2     641.8

Apartment Third of Higher Floor          46      1650  1696
                                       64.0    1632.0

All                                     298      7596  7894

  Cell Contents
      Count
      Expected count


Chi-Square Test

                    Chi-Square  DF  P-Value
          Pearson   9.589        2  0.008
 Likelihood Ratio   9.723        2  0.008
```

Example 12.6 Financial Support from Parents

● The report **"2019 Young Money Survey" (Ameritrade, June 2019)** summarized data from a survey of Generation Z, millennials, and parents with adult children. Table 12.4 summarizes responses from representative samples from each of these three groups. People in the sample of Generation Z and the sample of millennials were asked a question about the age they would consider it embarrassing to receive financial support from their parent(s). Possible responses were 20 to 29, 30 to 39, 40 to 49, and 50 or older. Each person in the sample of parents was asked about the age they thought it would be embarrassing for an adult child to still be receiving financial support from their parent(s).

Table 12.4 Observed and Expected Counts for Example 12.6

	Response				Row Marginal Total
	20 to 29	**30 to 39**	**40 to 49**	**50+**	
Generation Z	462 (512.34)	359 (326.56)	92 (72.95)	51 (52.15)	**964**
Millennials	451 (507.02)	349 (323.17)	82 (72.19)	72 (51.61)	**954**
Parents of Adult Children	590 (483.64)	250 (308.27)	40 (68.86)	30 (49.23)	**910**

Does there appear to be a difference in the response proportions for the three groups? The relevant hypotheses are

Understand the context ➤

H_0: The proportions for the age group response categories are the same for the three groups.

H_a: The age group response category proportions are not all the same for all three groups.

Significance level: $\alpha = 0.01$.

Formulate a plan ➤

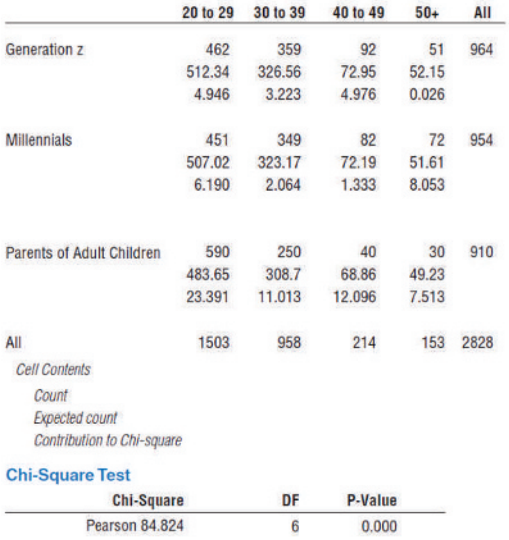
Test statistic: $$X^2 = \sum_{\text{all cells}} \frac{(\text{observed cell count} - \text{expected cell count})^2}{\text{expected cell count}}$$

Assumptions: Table 12.4 contains the calculated expected counts, all of which are greater than 5. The samples were independently selected and were said to be representative of the three populations of interest.

● **Data set available online**

Copyright 2025 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s). Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

Do the work ➤ Calculation: Minitab output follows. For each cell, the Minitab output includes the observed cell count, the expected cell count, and the value of $\frac{(\text{observed cell count} - \text{expected cell count})^2}{\text{expected cell count}}$ for that cell (this is the contribution to the X^2 statistic for this cell). From the output, $X^2 = 84.824$.



```
                          20 to 29  30 to 39  40 to 49     50+   All

Generation z                   462       359        92      51   964
                            512.34    326.56     72.95   52.15
                             4.946     3.223     4.976   0.026

Millennials                    451       349        82      72   954
                            507.02    323.17     72.19   51.61
                             6.190     2.064     1.333   8.053


Parents of Adult Children      590       250        40      30   910
                            483.65     308.7     68.86   49.23
                            23.391    11.013    12.096   7.513

All                           1503       958       214     153  2828
  Cell Contents
    Count
    Expected count
    Contribution to Chi-square

Chi-Square Test
               Chi-Square          DF      P-Value
           Pearson 84.824           6        0.000
```

P-value: From the Minitab output, P-value $= 0.000$.

Interpret the results ➤ Conclusion: Because the P-value is less than $\alpha = 0.01$, H_0 is rejected. The data indicate that the age group response proportions are not the same for the three populations. Comparing the observed and expected cell counts, you can see that more parents and fewer Generation Z and millennials chose the youngest age group response than would have been expected if there were no differences in the response distributions for the three populations. The two older age group responses (40 to 49 and 50+) were selected by more Generation Z and millennials and fewer parents than would have been expected if there were no differences in the response distributions.

Testing for Independence of Two Categorical Variables

A chi-square test can also be used to investigate the possibility of an association between two categorical variables in a single population. For example, television viewers in a particular city might be categorized with respect to both preferred network (ABC, CBS, NBC, or PBS) and favorite type of programming (comedy, drama, or information and news). The question of interest is often whether knowledge of the value of one variable provides any information about the value of the other variable—that is, are the two variables independent?

Continuing the example, suppose that those who favor ABC prefer the three types of programming in proportions 0.4, 0.5, and 0.1 and that these proportions are also correct for individuals favoring any of the other three networks. Then, learning an individual's preferred network provides no added information about that individual's favorite type of programming. The categorical variables *preferred network* and *favorite program type* would be independent.

To see how expected counts are obtained in this situation, recall from Chapter 6 that if two outcomes A and B are independent, then

$$P(A \text{ and } B) = P(A)P(B)$$

so the proportion of time that the two outcomes occur together in the long run is the product of the two individual long-run relative frequencies. Similarly, two categorical variables are independent in a population if, for each particular category of the first variable and each particular category of the second variable,

$$\begin{pmatrix}\text{proportion of individuals}\\ \text{in a particular category}\\ \text{combination}\end{pmatrix} = \begin{pmatrix}\text{proportion in}\\ \text{specified category}\\ \text{of first variable}\end{pmatrix}\cdot\begin{pmatrix}\text{proportion in}\\ \text{specified category}\\ \text{of second variable}\end{pmatrix}$$

This means that if 30% of all viewers prefer ABC and the proportions of program type preferences are as previously given, then, assuming that the two variables are independent, the proportion of individuals who both favor ABC and prefer comedy is $(0.3)(0.4) = 0.12$ (or 12%).

Multiplying the right-hand side of the expression above by the sample size gives the expected number of individuals in the sample who are in both specified categories of the two variables when the variables are independent. However, these expected counts cannot be calculated, because in reality, the individual population proportions are not known. The solution is to estimate each population proportion using the corresponding sample proportion:

$$\begin{pmatrix}\text{estimated expected number}\\ \text{in specified categories}\\ \text{of the two variables}\end{pmatrix} = (\text{sample size})\cdot\frac{\begin{pmatrix}\text{observed number}\\ \text{in category of}\\ \text{first variable}\end{pmatrix}}{\text{sample size}}\cdot\frac{\begin{pmatrix}\text{observed number}\\ \text{in category of}\\ \text{second variable}\end{pmatrix}}{\text{sample size}}$$

$$= \frac{\begin{pmatrix}\text{observed number in}\\ \text{category of first variable}\end{pmatrix}\cdot\begin{pmatrix}\text{observed number in}\\ \text{category of second variable}\end{pmatrix}}{\text{sample size}}$$

Suppose that the observed counts are displayed in a rectangular table in which rows correspond to the categories of the first variable and columns to the categories of the second variable. Then, the numerator in the expression for estimated expected counts is just the product of the row and column marginal totals. This is exactly how expected counts were calculated in the test for homogeneity of several populations, even though the reasoning that leads to the formula is different.

X^2 Test for Independence

Null Hypothesis: H_0: The two variables are independent (or equivalently, there is no association between the two variables).

Alternative Hypothesis: H_a: The two variables are not independent (or equivalently, there is an association between the two variables).

Test Statistic: $$X^2 = \sum_{\text{all cells}} \frac{(\text{observed cell count} - \text{expected cell count})^2}{\text{expected cell count}}$$

The expected cell counts are estimated (assuming H_0 is true) using the formula

$$\text{expected cell count} = \frac{(\text{row marginal total})(\text{column marginal total})}{\text{grand total}}$$

***P*-values:** When H_0 is true and the assumptions of the X^2 test for independence are satisfied, X^2 has approximately a chi-square distribution with

$$\text{df} = (\text{number of rows} - 1)(\text{number of columns} - 1)$$

The *P*-value associated with the calculated test statistic value is the area to the right of X^2 under the chi-square curve with the appropriate df. Upper-tail areas for chi-square distributions are found using technology or Appendix Table 8.

Assumptions:

1. The observed counts are from a sample that can be regarded as a random sample.
2. The sample size is large: All expected counts are at least 5. If some expected counts are less than 5, rows or columns of the table can be combined to achieve a table with satisfactory expected counts.

Example 12.7 A Pained Expression

● The paper **"Facial Expression of Pain in Elderly Adults with Dementia" (*Journal of Undergraduate Research* [2006])** examined the relationship between a nurse's assessment of a patient's facial expression and their self-reported level of pain. Data for 89 patients are summarized in Table 12.5.

Table 12.5 **Observed Counts for Example 12.7**

	Self-Report	
Facial Expression	**No Pain**	**Pain**
No Pain	17	40
Pain	3	29

The authors were interested in determining if there is evidence of a relationship between a facial expression that reflects pain and self-reported pain because patients with dementia do not always give a verbal indication that they are in pain.

Using a 0.05 significance level, the relevant hypotheses are:

Understand the context ➤

H_0: Facial expression and self-reported pain are independent.
H_a: Facial expression and self-reported pain are not independent.

Significance level: $\alpha = 0.05$.

Formulate a plan ➤

$$\text{Test statistic: } X^2 = \sum_{\text{all cells}} \frac{(\text{observed cell count} - \text{expected cell count})^2}{\text{expected cell count}}$$

Assumptions: Before you can check the assumptions you must first calculate the expected cell counts.

Cell		
Row	**Column**	**Expected Cell Count**
1	1	$\frac{(57)(20)}{89} = 12.81$
1	2	$\frac{(57)(69)}{89} = 44.19$
2	1	$\frac{(32)(20)}{89} = 7.19$
2	2	$\frac{(32)(69)}{89} = 24.81$

All expected cell counts are greater than 5. Although the participants in the study were not randomly selected, they were thought to be representative of the population of nursing home patients with dementia. The observed and expected counts are given together in Table 12.6.

Table 12.6 **Observed and Expected Counts for Example 12.7**

	Self-Report	
Facial Expression	**No Pain**	**Pain**
No Pain	17 (12.81)	40 (44.19)
Pain	3 (7.19)	29 (24.81)

Do the work ➤

$$\text{Calculation: } X^2 = \frac{(17 - 12.81)^2}{12.81} + \cdots + \frac{(29 - 24.81)^2}{24.81} = 4.92$$

P-value: The table has 2 rows and 2 columns, so df = $(2 - 1)(2 - 1) = 1$. The entry closest to 4.92 in the 1-df column of Appendix Table 8 is 5.02, so the approximate *P*-value for this test is

$$P\text{-value} \approx 0.025$$

● **Data set available online**

Copyright 2025 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s). Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

Interpret the results ➤

Conclusion: Since the P-value is less than α, you reject H_0 and conclude that there is convincing evidence that a nurse's assessment of facial expression and self-reported pain are not independent.

Example 12.8 Screen Use at Bedtime

● The National Sleep Foundation asked each person in a representative sample of 1082 adult Americans about bedtime screen use **("2022 Sleep in America Poll," sleepfoundation.org)**. Survey participants were classified into one of two screen use categories (often and not often). Each participant was also asked about race and ethnicity. Data for three race and ethnicity groups that are consistent with summary quantities given in the paper are given in Table 12.7. Expected cell counts (calculated under the assumption of no association between bedtime screen use and race and ethnicity) are also shown in Table 12.7.

Table 12.7 Observed and Expected Counts for Example 12.8

	Bedtime Screen Use	
Race/Ethnicity	**Often**	**Not Often**
Black	75 (80.79)	66 (60.21)
Hispanic	152 (130.07)	75 (96.93)
White	393 (409.13)	321 (304.87)

You can use these sample data to determine whether there is an association between race and ethnicity and bedtime screen use.

The X^2 test with a significance level of 0.01 will be used to test the relevant hypotheses:

Understand the context ➤

H_0: Race/ethnicity and bedtime screen use are independent.
H_a: Race/ethnicity and bedtime screen use are not independent.

Significance level: $\alpha = 0.01$.

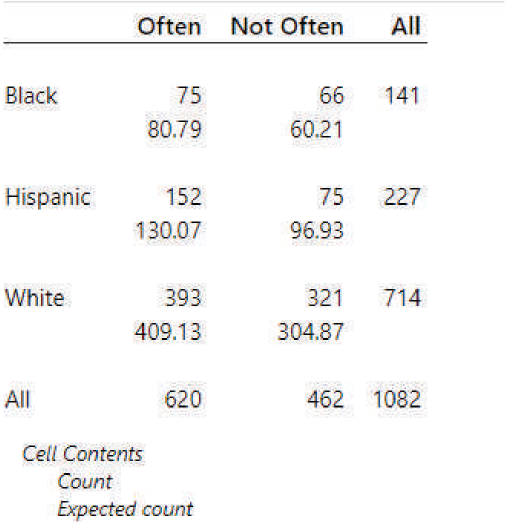
Formulate a plan ➤

$$\text{Test statistic: } X^2 = \sum_{\text{all cells}} \frac{(\text{observed cell count} - \text{expected cell count})^2}{\text{expected cell count}}$$

Assumptions: All expected cell counts are at least 5. Assuming that the sample is representative of adult Americans, the X^2 test can be used.

Do the work ➤

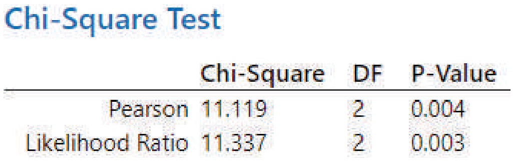
Calculation: Minitab output is shown. From the Minitab output, $X^2 = 11.119$.



```
             Often   Not Often    All

Black           75          66    141
             80.79       60.21

Hispanic       152          75    227
            130.07       96.93

White          393         321    714
            409.13      304.87

All            620         462   1082

  Cell Contents
     Count
     Expected count
```

● **Data set available online**

Copyright 2025 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s). Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.



```
Chi-Square Test

                  Chi-Square  DF  P-Value
         Pearson  11.119       2  0.004
Likelihood Ratio  11.337       2  0.003
```

P-value: From the Minitab output, *P*-value = 0.004.

Interpret the results ➤ Conclusion: Since the *P*-value is less than α, H_0 is rejected. There is convincing evidence that there is an association between race/ethnicity and bedtime screen use.

What Do You Think?

Based on a comparison of the observed and expected counts, how would you describe the association between race/ethnicity and bedtime screen use?

In some investigations, values of more than two categorical variables are recorded for each individual in the sample. For example, in addition to the variables *race and ethnicity* and *bedtime screen use*, the researchers in the study referenced in Example 12.8 might also have collected information on age. A number of interesting questions could then be explored: Are all three variables independent of one another? Is it possible that race and ethnicity and bedtime screen use are dependent but that the relationship between them does not depend on age?

The X^2 test procedure described in this section for analysis of bivariate categorical data can be extended for use with *multivariate categorical data*. Appropriate hypothesis tests can then be used to provide insight into the relationships between variables. However, the calculations required to determine expected cell counts and to calculate the value of X^2 are quite tedious, so they are seldom done without the aid of a computer. Most statistical software packages can perform this type of analysis.

Exercises 12.14–12.28

● **Data set available online**

12.14 A particular state university system has six campuses. On each campus, a random sample of students will be selected, and each student will be categorized with respect to political philosophy as liberal, moderate, or conservative. The null hypothesis of interest is that the proportion of students falling in each of these three categories is the same at all six campuses.

a. On how many degrees of freedom will the resulting X^2 test be based?

b. How does the answer in Part (a) change if there are seven campuses rather than six?

c. How does the answer in Part (a) change if there are four rather than three categories for political philosophy?

12.15 A random sample of 1000 registered voters in a certain county is selected, and each voter is categorized with respect to both educational level (four categories) and preferred candidate in an upcoming election for county supervisor (five possibilities). The hypothesis of interest is that educational level and preferred candidate are independent.

a. If $X^2 = 7.2$, what is the conclusion at significance level 0.10?

b. If there were only four candidates running for election, what would the conclusion be if $X^2 = 14.5$ and $\alpha = 0.05$?

12.16 Explain the difference between situations that would lead to a chi-square test for homogeneity and those that would lead to a chi-square test for independence.

12.17 Give an example of a situation where it would be appropriate to use a chi-square test of homogeneity. Describe the populations that would be sampled and the variable that would be recorded.

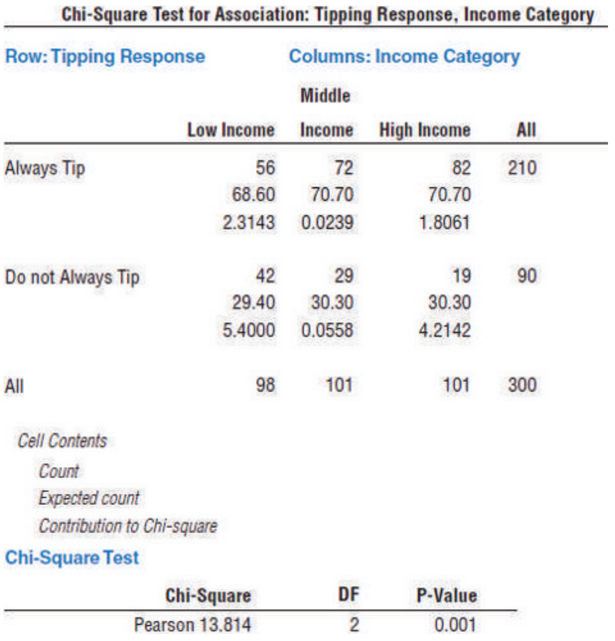
12.18 Give an example of a situation where it would be appropriate to use a chi-square test of independence. Describe the population that would be sampled and the two variables that would be recorded.

Copyright 2025 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s). Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

12.19 ● The online article **"COVID-19 Has Not Made Americans Better Tippers" (creditcards.com, July 12, 2021)** claims that "tipping is strongly related to income." Suppose that each person in a random sample of 300 adults was asked if they always tip when eating at a restaurant and was also classified into one of three income groups (low, middle, high). The data in the accompanying table are consistent with summary statistics given in the article).

	Income Group		
	Low Less Than $40,000 per Year	**Middle Between $40,000 and $80,00 per Year**	**High More Than $80,000 per Year**
Always Tip	56	72	82
Do Not Always Tip	42	29	19

a. Use the accompanying Minitab output to carry out an appropriate hypothesis test to determine if there is convincing evidence to support the claim made in the article. Use a significance level of $\alpha = 0.05$.



```
Chi-Square Test for Association: Tipping Response, Income Category

Row: Tipping Response          Columns: Income Category

                                Middle
                   Low Income   Income   High Income   All

Always Tip                 56       72            82   210
                        68.60    70.70         70.70
                       2.3143   0.0239        1.8061

Do not Always Tip          42       29            19    90
                        29.40    30.30         30.30
                       5.4000   0.0558        4.2142

All                        98      101           101   300

  Cell Contents
    Count
    Expected count
    Contribution to Chi-square
Chi-Square Test

               Chi-Square    DF   P-Value
          Pearson 13.814      2     0.001
```

b. Write a few sentences describing how tipping response is related to income. Hint: Focus on cells that have large values of

$$\frac{(\text{observed cell count} - \text{expected cell count})^2}{\text{expected cell count}}$$

12.20 ● Each person in a representative sample of 445 college students age 18 to 24 was classified according to age and to the response to the following question: "How often have you used a credit card to buy items knowing you wouldn't have money to pay the bill when it arrived?" Possible responses were never, rarely, sometimes, or frequently **("Majoring in Money: How American College Students Manage Their Finances," June 28, 2016, salliemae.newshq.businesswire.com)**. The responses are summarized in the accompanying table. Do these data provide evidence that there is an association between age group and the response to the question? Test the relevant hypotheses using $\alpha = 0.01$.

	Age 18to 20	**Age 21to 22**	**Age 23 to 24**
Never	72	62	29
Rarely	36	34	32
Sometimes	30	42	40
Frequently	12	24	32

12.21 ● The paper **"Understanding the Deterrent Impact Formal and Informal Sanctions Have on Illegal Smartphone Use While Driving" (*Accident Analysis and Prevention* [2020]: 105706)** describes a survey of a large sample of drivers. Each driver was asked about their frequency of illegal smartphone use while driving and was also classified according to age. The resulting data are summarized in the accompanying table. For purposes of this exercise, assume that it is reasonable to regard this sample as representative of the population of drivers. Carry out a test to determine if it is reasonable to conclude that there is an association between age group and frequency of illegal smartphone use while driving.

		Age			
		18to 25	**26 to 39**	**40 to 59**	**60+**
Frequency of Illegal Smartphone Use While Driving	**Never**	155	288	635	668
	Half the Time or Less	204	256	278	86
	More Than Half the Time	102	54	43	5

12.22 ● The report **"Consumer Revolving Credit and Debt Over the Life Cycle and Business Cycle"** describes a study conducted by the **Federal Reserve Bank of Boston (bostonfed.org, October 2015)**. Data consistent with summary values given in the report are summarized in the accompanying table. Suppose that these data resulted from a random sample of 800 adult Americans age 20 to 39 years old who have at least one credit card. Each person in the sample was classified according to age (with possible categories of 20 to 24 years, 25 to 29 years, 30 to 34 years, and 35 to 39 years). The people in the sample were also classified according to whether or not they pay the full balance on their credit cards each month or sometimes carry over a balance from month to month.

	Pay Full Balance Each Month	Carry Balance from Month to Month
Age 20 to 24 years	75	95
Age 25 to 29 years	74	126
Age 30 to 34 years	75	145
Age 35 to 39 years	67	143

a. To investigate if whether or not people pay their balance in full each month is related to age, which chi-square test (homogeneity or independence) would be the appropriate test? Explain your choice.

b. Carry out an appropriate test to determine if these data provide convincing evidence that whether or not people pay their balance in full each month is related to age.

c. To what population would it be reasonable to generalize the conclusion from the test in Part (b)?

12.23 ● Artificial intelligence algorithms are now capable of generating facial images that are difficult to distinguish from photos of real faces. The authors of the paper **"AI-Synthesized Faces are Indistinguishable from Real Faces and More Trustworthy" (*PNAS* [2022]: e2120481119)** collected data to investigate the claim that synthetic faces are judged to be more trustworthy than real faces. Suppose that a sample of people were asked to rate the trustworthiness of a photo of a real person and a different sample of people were asked to rate a photo of a synthesized face that was similar to the real face in terms of gender, race, age, and weight. Also suppose that the trustworthiness response could be one of seven categories from 1 (least trustworthy) to 7 (most trustworthy). Response data are given in the accompanying table. (Note that this is not the same as the experiment described in the paper—it has been simplified for purposes of this exercise. But the data are consistent with the findings of the paper.) Do the data in the accompanying table provide convincing evidence that the trustworthiness response distribution is not the same for the real face and the synthesized face? Use $\alpha = 0.01$ for your test.

	Trustworthiness Response				
	1 or 2	3	4	5	6 or 7
Real Face	18	69	166	138	4
Synthesized Face	4	40	145	171	17

12.24 ● Ipsos, a market research company, surveyed a representative sample of 1005 adult Americans. Each person in the sample was asked about the region of the country where they lived and whether they had a tattoo or multiple tattoos on their body **("More Americans Have Tattoos Today Than Seven Years Ago," ipsos.com, August 29, 2019)**. The responses are summarized in the accompanying table. Do these data provide convincing evidence of an association between regions of the country and tattoo category? Use $\alpha = 0.01$.

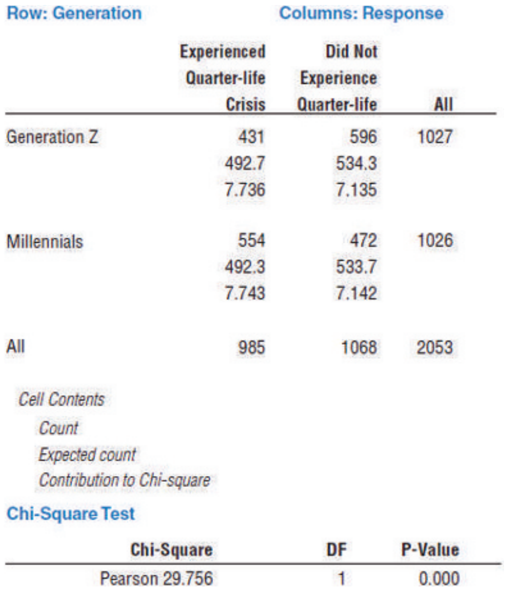
	Region			
	Northeast	Midwest	South	West
One Tattoo	27	33	42	30
Multiple Tattoos	27	36	66	41
No Tattoos	125	143	269	166

12.25 ● The report **"2019 Young Money Survey" Ameritrade, June 2019)** summarized data from a survey of Generation Z and millennials. Each person in a representative sample of Generation Z and a representative sample of millennials was asked if they had experienced a "quarter-life crisis." A quarter-life crisis involves anxiety about the direction and quality of life experienced in a person's early 20s to mid-30s. For the Generation Z sample, 431 of the 1027 people surveyed said they had experienced a quarter-life crisis, and 554 of the 1026 millennials surveyed said they had experienced a quarter-life crisis. The responses are summarized in the accompanying table.

	Experienced Quarter-Life Crisis	Did Not Experience Quarter-Life Crisis
Generation Z	431	596
Millennials	554	472

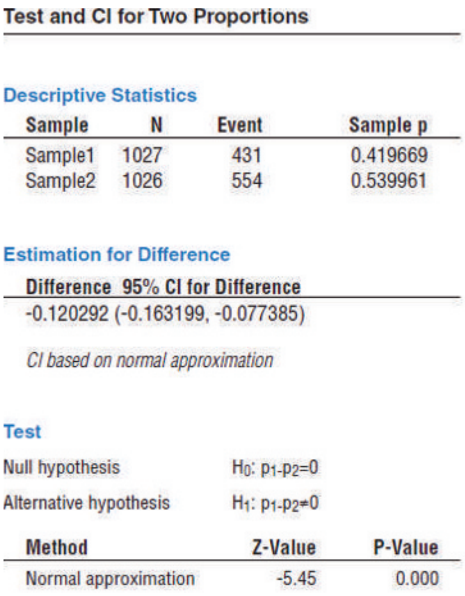
a. Is there evidence that the proportions who would fall into each of the two response categories (did experience quarter-life crisis and did not experience quarter-life crisis) are not the same for Generation Z and millennials? Use a X^2 test to test the relevant hypotheses with a significance level of $\alpha = 0.05$.

b. Are your calculations and conclusions from Part (a) consistent with the accompanying Minitab output? Explain.



```
Row: Generation                    Columns: Response

                  Experienced      Did Not
                  Quarter-life  Experience
                        Crisis  Quarter-life    All
Generation Z               431          596    1027
                         492.7        534.3
                         7.736        7.135

Millennials                554          472    1026
                         492.3        533.7
                         7.743        7.142

All                        985         1068    2053

 Cell Contents
   Count
   Expected count
   Contribution to Chi-square

Chi-Square Test

            Chi-Square        DF      P-Value
          Pearson 29.756       1        0.000
```

c. Because the response variable in this exercise has only two categories (experienced quarter-life crisis and did not experience quarter-life crisis), you could have also answered the question posed by carrying out a large-sample z test of H_0: $p_1 - p_2 = 0$ versus H_a: $p_1 - p_2 \neq 0$, where p_1 is the proportion who experienced a quarter-life crisis for Generation Z and p_2 is the proportion who experienced a quarter-life crisis for millennials. Minitab output from the large-sample z test is shown. Using a significance level of $\alpha = 0.05$, does the large-sample z test lead to the same conclusion as in Part (a)?



```
Test and CI for Two Proportions

Descriptive Statistics
  Sample     N     Event     Sample p
  Sample1   1027    431      0.419669
  Sample2   1026    554      0.539961

Estimation for Difference
  Difference  95% CI for Difference
  -0.120292 (-0.163199, -0.077385)

  CI based on normal approximation

Test
Null hypothesis          H0: p1-p2=0
Alternative hypothesis   H1: p1-p2≠0

  Method                  Z-Value    P-Value
  Normal approximation     -5.45      0.000
```

d. How do the P-values from the tests in Parts (a) and (c) compare? Does this surprise you? Explain.

12.26 ● The infographic **"Smartphone Usage and Penetration in 2021" (smsglobal, August 17, 2021)** provided the following information on smartphone ownership for representative samples of adults in several different countries.

Country	Percent Who Own a Smartphone	Percent Who Do Not Own a Smartphone
United States	82%	18%
Spain	74%	26%
Italy	76%	24%
Germany	78%	22%

a. Suppose the sample sizes were 1000 for the United States and for Germany and 500 for Spain and Italy. Complete the following two-way table by entering the observed counts.

Country	Own a Smartphone	Do Not Own a Smartphone
United States		
Spain		
Italy		
Germany		

b. Carry out a hypothesis test to determine if there is convincing evidence that the smartphone ownership proportions are not the same for all four countries. Use a significance level of $\alpha = 0.05$.

12.27 ● The online article **"Nation's Teachers More Diverse But Still Lag Racial, Ethnic Makeup of Students" (U.S. Census Bureau, January 5, 2022)** used data from the city of Houston to investigate the gap in teacher/student diversity. Suppose that each person in a representative sample of 400 public school students and a representative sample of 400 public school teachers was asked about their race and ethnicity. The data in the accompanying table are consistent with summary values for Houston that appeared in article. Is there convincing evidence that the race/ethnicity proportions differ for students and teachers?

Race/Ethnicity	Students	Teachers
White Alone	75	234
Hispanic or Latino (Any Race)	264	79
Black or African American Alone	103	129
Asian Alone	25	37
American Indian/Alaska Native Alone	5	6
Native Hawaiian/Pacific Islander Alone	6	4
Two or More Races	22	11

12.28 ● The report **"Teens, Social Media and Technology 2022" (Pew Research Center, 2022)** surveyed a representative sample of teens age 13 to 17. Each survey participant was asked about the amount of time that they spend on social media, with possible responses of too much, about right, or too little. Participants were also classified into location categories based on whether they lived in an urban, suburban, or rural area, resulting in the data given in the accompanying table. Do these data provide convincing evidence that there is an association between the response to the social media question and location?

	Too Much	About Right	Too Little
Urban	121	126	40
Suburban	239	414	49
Rural	116	183	22

Section 12.3 Interpreting and Communicating the Results of Statistical Analyses

Many studies, particularly those in the social sciences, result in categorical data. The questions of interest in such studies often lead to an analysis that involves using a chi-square test.

Communicating the Results of Statistical Analyses

Three different chi-square tests were introduced in this chapter—the goodness-of-fit test, the test for homogeneity, and the test for independence. They are used in different settings and to answer different questions. When summarizing the results of a chi-square test, be sure to indicate which chi-square test was performed. One way to do this is to be clear about how the data were collected and the nature of the hypotheses being tested.

It is also a good idea to include a table of observed and expected counts in addition to reporting the value of the test statistic and the *P*-value. And finally, make sure to give a conclusion in context, and word the conclusion appropriately for the type of test conducted. For example, don't use terms such as *independence* and *association* to describe the conclusion if the test performed was a test for homogeneity.

Interpreting the Results of Statistical Analyses

As with the other hypothesis tests considered, it is common to find the result of a chi-square test summarized by giving the value of the chi-square test statistic and an associated *P*-value. Because categorical data can be summarized compactly in frequency tables, the data are often given in the article (unlike data for numerical variables, which are rarely given).

What to Look for in Published Data

Here are some questions to consider when reading an article that includes the results of a chi-square test:

- Are the variables of interest categorical rather than numerical?
- Are the data given in the article in the form of a frequency table?
- If a two-way frequency table is involved, is the question of interest one of homogeneity or one of independence?
- What null hypothesis is being tested? Are the results of the analysis reported in the correctly in context?
- Are sample sizes large enough to make use of a chi-square test reasonable? (Are all expected counts at least 5?)
- What is the value of the test statistic? Is the associated *P*-value given? Should the null hypothesis be rejected?
- Are the conclusions drawn consistent with the results of the test?
- How different are the observed and expected counts? Does the result have practical significance as well as statistical significance?

The paper **"Association of Flavor Perception With Blue vs Purple Cigar Packaging Among US Adults" (*JAMA Network Open* [2023]: e2254003)** carried out an experiment to investigate whether the color of packaging can influence consumer perception of a product. Study participants were randomly assigned to one of two treatment groups. One group (the blue group) viewed a cigar in blue packaging and the other group (the purple group) viewed the same cigar in purple packaging. The reported sample sizes were 1319 for the blue group and 1222 for the purple group. After viewing the cigars, participants were asked several questions. They were first asked about their own cigar use, with possible responses of never smoked a cigar, have ever smoked a cigar but not a current smoker, and current cigar smoker. They were also asked whether they thought that the cigar they viewed was

Copyright 2025 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s). Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

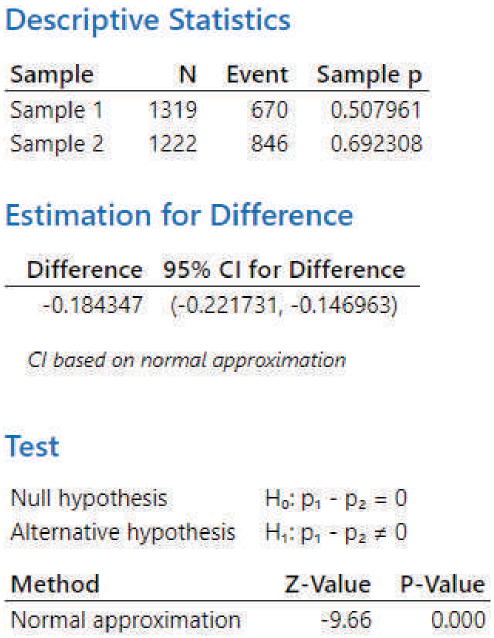
flavored. If they thought the cigar was flavored, they had the opportunity to say what they thought the flavor was.

The authors stated that

> More participants randomized to purple vs blue believed that the product was flavored (69.2% vs 50.8%.)

While the authors did not state that this observed difference in proportions was significant, this could be tested using either a large sample test of treatment proportions (a chi-square test for homogeneity could be used to test to see if the proportions are different, but because the stated conclusion suggests use of an upper-tail test, the large-sample test for proportions would be preferred here). The sample sizes are large and participants were randomly assigned to treatments, so the large-sample z test for proportions is an appropriate choice.

Minitab output is shown here, and supports that the percentage of people who believe that the cigar is flavored when it is in purple packaging is in fact significantly greater than this proportion for blue packaging. The confidence interval provides an estimate of how much greater—between 14 and 22 percentage points.



```
Descriptive Statistics

Sample       N  Event  Sample p
Sample 1  1319    670  0.507961
Sample 2  1222    846  0.692308

Estimation for Difference

  Difference  95% CI for Difference
   -0.184347  (-0.221731, -0.146963)

  CI based on normal approximation

Test

Null hypothesis         H₀: p₁ - p₂ = 0
Alternative hypothesis  H₁: p₁ - p₂ ≠ 0

Method                Z-Value  P-Value
Normal approximation    -9.66    0.000
```

The authors also concluded that previous cigar smoking was

> associated with believing that the cigar blend was flavored. Significant differences were observed for perceived flavors by color with blue evoking vanilla, blueberry, or mint/menthol and purple evoking grape, wine, or berry/cherry.

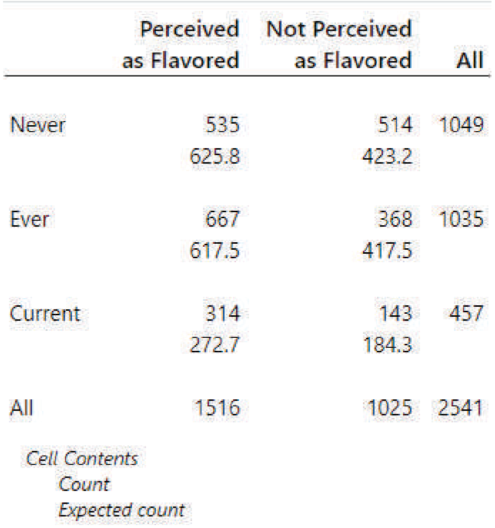
The first part of this statement can be justified by carrying out a chi-square test of homogeneity. Data from the paper are summarized in the following table:

	Perceived as Flavored	**Not Perceived as Flavored**
Never Smoked	535	514
Have Smoked But Not Current Smoker	667	368
Current Smoker	314	143

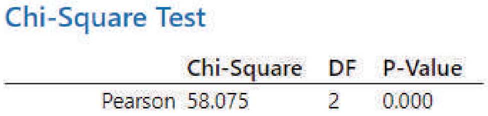
To determine if previous cigar smoking is associated with whether the cigar viewed is perceived as flavored, the following hypotheses could be tested:

H_0: there is an association between previous cigar smoking and flavor preception
H_a: there is not an association between previous cigar smoking and flavor preception

From the Minitab output below, $X^2 = 58.075$, and P-value $= 0.000$, so there is a significant association. Comparing observed and expected counts shows that more people who never smoked cigars thought that the cigar was flavored and fewer people who were current cigar smokers or who had smoked cigars in the past thought that the cigar was flavored than would have been expected if there were no association



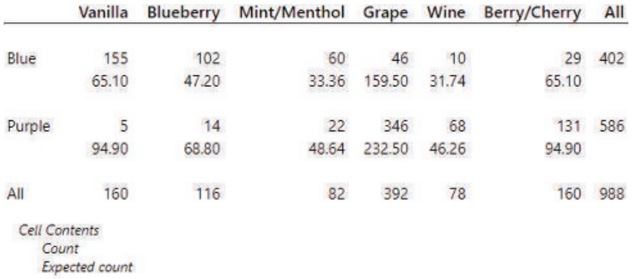
```
              Perceived   Not Perceived
            as Flavored     as Flavored   All

Never               535             514  1049
                  625.8           423.2

Ever                667             368  1035
                  617.5           417.5

Current             314             143   457
                  272.7           184.3

All                1516            1025  2541

  Cell Contents
      Count
      Expected count
```



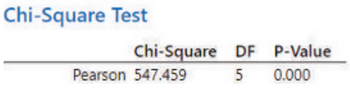
```
Chi-Square Test

             Chi-Square   DF   P-Value
      Pearson    58.075    2     0.000
```

The second part of the statement above was that significant differences were observed for perceived flavors by color with blue evoking vanilla, blueberry, or mint/menthol and purple evoking grape, wine, or berry/cherry. People who said that they thought the cigar was flavored had the opportunity to state what they thought the flavor was. The data for those who chose to give an opinion of the flavor was used to generate the accompanying Minitab output. Notice that all expected counts are greater than 5 and study participants were randomly assigned to the packaging color groups, so the chi-square test of homogeneity is an appropriate method to use to test the hypotheses

H_0: the proportions choosing each of the flavor choice categories is the same for blue and purple packaging
H_a: the proportions choosing each of the flavor choice categories is not the same for blue and purple packaging



```
        Vanilla  Blueberry  Mint/Menthol   Grape   Wine  Berry/Cherry  All

Blue        155        102            60      46     10            29  402
          65.10      47.20         33.36  159.50  31.74         65.10

Purple        5         14            22     346     68           131  586
          94.90      68.80         48.64  232.50  46.26         94.90

All         160        116            82     392     78           160  988

  Cell Contents
      Count
      Expected count
```



```
Chi-Square Test

            Chi-Square  DF  P-Value
    Pearson    547.459   5    0.000
```

Copyright 2025 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s). Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

From the Minitab output, $X^2 = 547.459$, and P-value $= 0.000$, which would lead to rejection of the null hypothesis. The large differences between observed and expected counts support the authors' interpretation about the differences in flavor perceptions for the two packaging colors.

A Word to the Wise: Cautions and Limitations

Be sure to keep the following in mind when analyzing categorical data using one of the chi-square tests presented in this chapter:

1. Don't confuse tests for homogeneity with tests for independence. The hypotheses and conclusions are different for the two types of test. Tests for homogeneity are used when the individuals in each of two or more independent samples are classified according to a single categorical variable. Tests for independence are used when individuals in a *single* sample are classified according to two categorical variables.
2. As was the case for the hypothesis tests of earlier chapters, remember that you can never say you have strong support for the null hypothesis. For example, if you do not reject the null hypothesis in a chi-square test for independence, you cannot conclude that there is convincing evidence that the variables are independent. You can only say that you were not convinced that there is an association between the variables.
3. Be sure that the assumptions for the chi-square test are reasonable. P-values based on the chi-square distribution are only approximate, and if the large sample conditions are not met, the actual P-value may be quite different from the approximate one based on the chi-square distribution. This can sometimes lead to incorrect conclusions. Also, for the chi-square test of homogeneity, the assumption of *independent* samples is particularly important.
4. Don't jump to conclusions about causation. Just as a strong correlation between two numerical variables does not mean that there is a cause-and-effect relationship between them, an association between two categorical variables does not imply a causal relationship.

Exercises 12.29–12.30

12.29 The authors of the paper **"Fruit, Vegetable, and Sugar-Sweetened Beverage Intake Among Young Children, by State—United States, 2021" (National Library of Medicine, February 17, 2023)** analyzed data from the 2021 National Survey of Children's Health. In this survey, parents of children age 1 to 5 years reported how often their children consumed fruits, vegetables, and sugar-sweetened beverages. The following statement appeared in the paper:

> Children aged 1 year were more likely than were older children to eat either a daily fruit or a daily vegetable during the preceding week and were less likely to drink a sugar-sweetened beverage (chi-square $p < 0.05$).

This statement summarizes the results of three different X^2 tests of independence using data in two-way tables.

The first part of the given statement says that 1-year-old children were more likely to eat fruit daily than older children. This statement could be based on a X^2 test of independence using data in a two-way table where rows corresponded to age (with five categories corresponding to ages 1, 2, 3, 4, and 5). Columns in the table correspond to two consumption categories (daily and less than daily).

a. What are the hypotheses for this test? What is the number of degrees of freedom associated with this test?

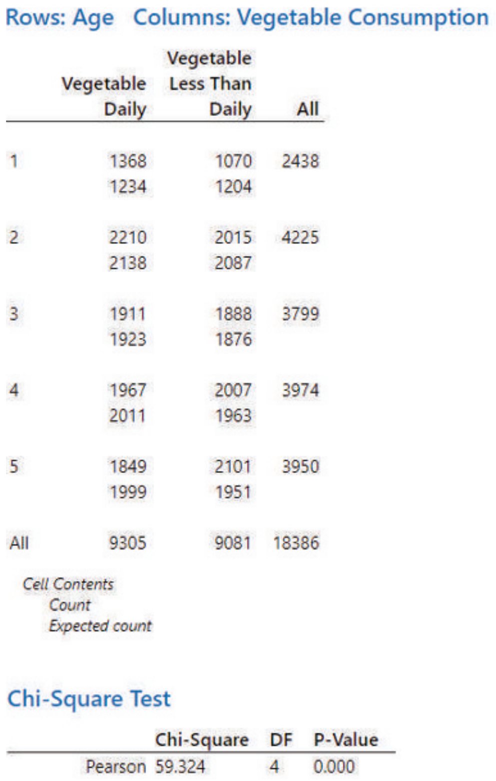
b. Using data that appeared in the paper, the calculated value of the X^2 statistic is 93.7. What is the P-value associated with this test? Using a significance level of 0.05, what conclusion would you reach?

c. Would a different conclusion have been reached if a significance level of 0.01 had been used for the test?

d. The paper also included data on the vegetable consumption, which is summarized in the following table.

Age	Vegetable Daily	Vegetable Less Than Daily
1	1368	1070
2	2210	2015
3	1911	1888
4	1967	2007
5	1849	2101

Use the Minitab output that follows to carry out a X^2 test of independence. Are the conclusion of this test and the differences in observed and expected counts consistent with the statement in the paper that says that 1-year-old children were more likely to eat vegetables daily than older children?



```
Rows: Age   Columns: Vegetable Consumption

                        Vegetable
            Vegetable   Less Than
                Daily       Daily     All
-----------------------------------------

1                1368        1070    2438
                 1234        1204

2                2210        2015    4225
                 2138        2087

3                1911        1888    3799
                 1923        1876

4                1967        2007    3974
                 2011        1963

5                1849        2101    3950
                 1999        1951

All              9305        9081   18386

  Cell Contents
     Count
     Expected count


Chi-Square Test

              Chi-Square   DF   P-Value
---------------------------------------
     Pearson  59.324        4   0.000
```

12.30 Using data from a national survey, the authors of the paper **"What Do Happy People Do?" (*Social Indicators Research* [2008]: 565–571)** concluded that there was convincing evidence of an association between amount of time spent watching television and whether or not a person reported that they were happy. They observed that unhappy people tended to watch more television. The authors write:

> This could lead us to two possible interpretations:
>
> 1. Television viewing is a pleasurable enough activity with no lasting benefit, and it pushes aside time spent in other activities—ones that might be less immediately pleasurable, but that would provide long-term benefits in one's condition. In other words, television does cause people to be less happy.
> 2. Television is a refuge for people who are already unhappy. TV is not judgmental nor difficult, so people with few social skills or resources for other activities can engage in it. Furthermore, chronic unhappiness can be socially and personally debilitating and can interfere with work and most social and personal activities, but even the unhappiest people can click a remote and be passively entertained by a TV. In other words, the causal order is reversed for people who watch television; unhappiness leads to television viewing.

Using only data from this study, is it possible to determine which of these two conclusions is correct? If so, which conclusion is correct and why? If not, explain why it is not possible to decide which conclusion is correct based on the study data.

Chapter Activities

Activity 12.1 Pick a Number, Any Number . . .

The purpose of this activity is to use a chi-square test to investigate whether people are good at generating random numbers.

Background: There is evidence to suggest that human beings are not very good random number generators. In this activity, you will investigate this phenomenon by collecting and analyzing a set of human-generated "random" digits.

For this activity, work in a group with four or five other students.

1. Each member of the group should complete this step individually. Ask 25 different people to pick a digit from 0 to 9 at random. Record the responses.
2. Combine the responses you collected with those of the other members of your group to form a single

Copyright 2025 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s). Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

sample. Summarize the resulting data in a one-way frequency table.

3. If people are adept at picking digits at random, what would you expect for the proportion of the responses in the sample that were 0? that were 1?
4. State a null hypothesis and an alternative hypothesis that could be tested to determine whether there is evidence that the 10 digits from 0 to 9 are not selected an equal proportion of the time when people are asked to pick a digit at random.
5. Carry out the appropriate hypothesis test, and write a few sentences indicating whether or not the data support the theory that people are not good random number generators.

Activity 12.2 Color and Perceived Taste

The purpose of this activity is to provide an opportunity to review how to design and carry out an experiment that results in data that can be analyzed using a chi-square test.

Background: Does the color of a food or beverage affect the way people perceive its taste? In this activity you will conduct an experiment to investigate this question and analyze the resulting data using a chi-square test.

You will need to recruit at least 30 subjects for this experiment, so it is advisable to work in a large group (perhaps even the entire class) to complete this activity.

Subjects for the experiment will be assigned at random to one of two groups. Each subject will be asked to taste a sample of gelatin (for example, Jell-O) and rate the taste as not very good, acceptable, or very good. Subjects assigned to the first group will be asked to taste and rate a cube of lemon-flavored gelatin. Subjects in the second group will be asked to taste and rate a cube of lemon-flavored gelatin that has been colored an unappealing color by adding food coloring to the gelatin mix before the gelatin sets.

Note: You may choose to use something other than gelatin, such as lemonade. Any food or beverage whose color can be altered using food coloring can be used. You can experiment with the food colors to obtain a color that you think is particularly unappealing!

1. As a class, develop a plan for collecting the data. How will subjects be recruited? How will they be assigned to one of the two treatment groups (unaltered color, altered color)? What extraneous variables will be directly controlled, and how will you control them?
2. After the class is satisfied with the data collection plan, assign members of the class to prepare the gelatin to be used in the experiment.
3. Carry out the experiment, and summarize the resulting data in a two-way table like the one shown:

	Taste Rating		
	Not Very Good	**Acceptable**	**Very Good**
Unaltered Color			
Altered Color			

4. The two-way table summarizes data from two independent samples (as long as subjects were assigned *at random* to the two treatments, the samples are independent). Carry out an appropriate test to determine whether the proportion for each of the three taste rating categories is the same when the color is altered as when the color is not altered.

Summary Key Concepts and Formulas

Term or Formula	Comment
One-way frequency table	A compact way of summarizing data on a categorical variable. It gives the number of times each of the possible categories in the data set occurs (the frequencies).
$X^2 = \sum_{\text{all cell}} \frac{(\text{observed cell count} - \text{expected cell count})^2}{\text{expected cell count}}$	A statistic used to provide a comparison between observed counts and those expected when a given hypothesis is true. When none of the expected counts are too small, X^2 has approximately a chi-square distribution.
X^2 goodness-of-fit test	A hypothesis test performed to determine whether population category proportions are different from those specified by a given null hypothesis.
Two-way frequency table (contingency table)	A rectangular table used to summarize a categorical data set. Two-way tables are used to compare several populations on the basis of a categorical variable or to determine if an association exists between two categorical variables.
X^2 test for homogeneity	The hypothesis test performed to determine whether category proportions are the same for two or more populations or treatments.
X^2 test for independence	The hypothesis test performed to determine whether an association exists between two categorical variables.

Chapter Review Exercises 12.31–12.36

● **Data set available online**

12.31 ● The International Survey of Pet Owners and Veterinarians **(Human Animal Bond Research Institute, January 16, 2022)** asked each person in a representative sample of cat owners and a representative sample of dog owners to respond to the following question: Which of the following best describes how you think about your pet? Possible responses were a family member, like a child, a friend, a companion, and just a pet. For purposes of this exercise, suppose that the two sample sizes were each 1000 (the actual sample sizes were much larger). Data consistent with summary statistics from the survey are given in the accompanying table. Based on these data, it is reasonable to conclude that the distribution of responses is not the same for cat owners and dog owners? Test the appropriate hypotheses using $\alpha = 0.05$.

	Family Member	Like a Child	Friend	Companion	Just a Pet
Cat Owners	450	350	90	90	20
Dog Owners	460	320	80	100	40

12.32 ● The 2020 Consumer Financial Literacy Survey summarized data from a representative sample of adult Americans. Survey participants were asked if they felt confident in their most recent big financial decision and were also classified into groups by age. Data consistent with summary percentages given in the report are summarized in the accompanying table.

	Confident in Decision	Not Confident in Decision
Age 18to 34	260	50
Age 35 to 44	141	29
Age 45 to 54	174	26
Age 55 to 64	160	20
Age 65 and Older	133	7

a. Do these data provide convincing evidence of an association between age group and confidence in most recent big financial decision? Use $\alpha = 0.05$.

b. Write a few sentences commenting on the nature of any association between age group and confidence in most recent big financial decision.

12.33 ● The article **"Do Politics Matter When It Comes to Loving Cats or Dogs?" (psychologytoday.com, June 25, 2013)** summarizes data from a survey of 603 American adults conducted by Public Policy Polling. One question in the survey was "Based on their personalities, do you think a dog or a cat would make a better president of the United States?" Survey participants were also asked if they would describe themselves as liberal or conservative. Use the data in the accompanying table to determine if there is convincing evidence of an association between political philosophy (liberal or conservative) and the response to the question about whether a dog or a cat would make a better president (possible responses were dog, cat, and undecided). Use a significance level of 0.05 for your test.

	Liberal	Conservative
Dog Would Make a Better President	64	171
Cat Would Make a Better President	69	36
Undecided	114	149

12.34 Data from a representative sample of American adults is summarized in the report **"Americans and Privacy: Concerned, Confused and Feeling Lack of Control Over Their Personal Information" (Pew Research Center, 2019)**. Survey participants were classified according to their reported political affiliation into two groups: Democrats (including independents who lean toward the Democratic party) and Republicans (including independents who lean toward the Republican party). They also responded to a question asking if they think there should be less, about the same, or more regulation of what companies can do with their customers' personal information. Data from the report was used to reach the following conclusion:

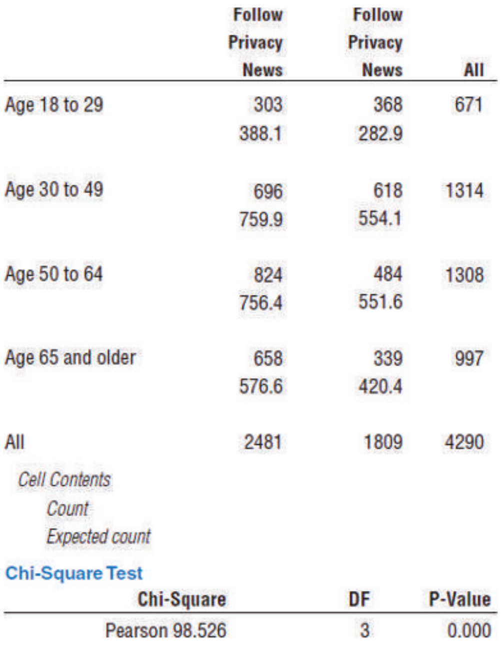
> There is evidence at the 0.05 level of significance that there is an association between political affiliation and response to the question about government regulation about use of personal information ($X^2 = 47.5$, P-value $= 0.000$). Democrats (including independents who lean toward the Democratic party) are more likely than Republicans (including independents who lean toward the Republican party) to believe there should be more government regulation of what companies can do with their customers' personal information (81% vs. 70%).

This conclusion was based on carrying out a chi-square test of independence using a two-way table where rows represented political affiliation category and columns represented the response to the question about government regulation.

a. What hypotheses were tested? What is the number of degrees of freedom associated with the given value of the X^2 statistic?

b. Would the stated conclusion have been different if a significance level of 0.01 had been used? Explain.

12.35 The paper referenced in the previous exercise also included data on age group and whether or not the survey participant said that they follow news about privacy related issues. Use the accompanying Minitab output to carry out a test to determine if there is evidence that there is an association between age and whether or not news about privacy related issues is followed?



```
                   Follow    Follow
                   Privacy   Privacy
                   News      News      All
Age 18 to 29         303       368     671
                   388.1     282.9

Age 30 to 49         696       618    1314
                   759.9     554.1

Age 50 to 64         824       484    1308
                   756.4     551.6

Age 65 and older     658       339     997
                   576.6     420.4

All                 2481      1809    4290

  Cell Contents
     Count
     Expected count

Chi-Square Test
            Chi-Square        DF    P-Value
       Pearson 98.526          3      0.000
```

12.36 ● The report **"Education Pays 2019" (The College Board, 2019)** provided information on education level and earnings for a sample of adult Americans who are employed full-time. Data consistent with summary percentages given in the report are summarized in the accompanying table. Suppose this data resulted from a representative sample of 1000 working adults whose highest level of education was either a high school diploma, an associate degree, or a bachelor's degree. Each person in the sample was classified according to education level (high school diploma, associate degree, or bachelor's degree) and yearly income (with possible categories of less than $20,000, $20,000 to $39,999, $40,000 to $59,999, and $60,000 or more). Is there evidence of an association between income category and education level? Test the appropriate hypotheses using a 0.05 significance level.

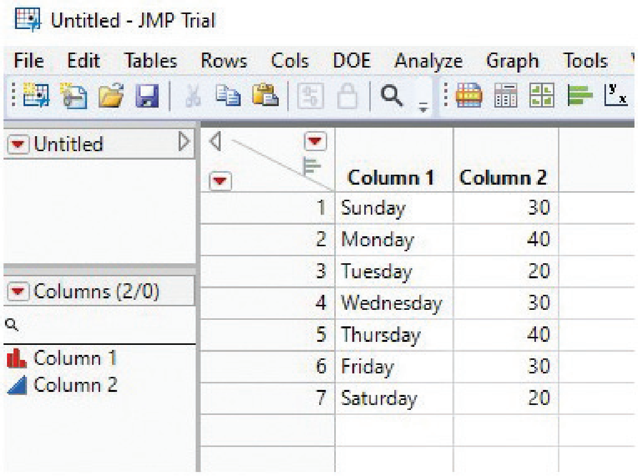
	Income Level			
	Less than $20,000	**$20,000 to $39,999**	**$40,000 to $59,999**	**$60,000 or more**
High School Diploma	30	152	103	95
Associate Degree	7	49	56	68
Bachelor's Degree	9	66	97	268

Technology Notes

X^2 Goodness-of-Fit Test

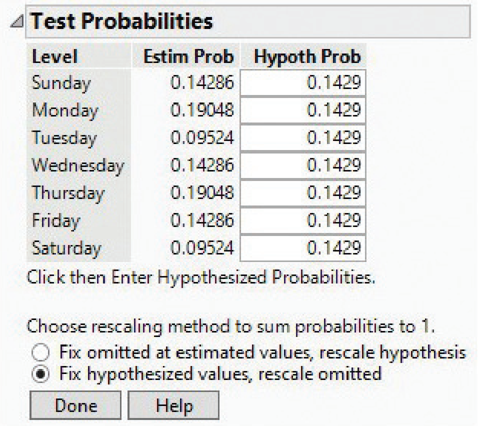
JMP

1. Enter each category name into the first column
2. Enter the unt for each category into the second column



	Column 1	Column 2
1	Sunday	30
2	Monday	40
3	Tuesday	20
4	Wednesday	30
5	Thursday	40
6	Friday	30
7	Saturday	20

3. Click **Analyze** then select **Distribution**
4. Click and drag the name for the first column from the box under **Select Columns** to the box next to **Y, Columns**
5. Click and drag the name for the second column from the box under **Select Columns** to the box next to **Freq**
6. Click **OK**
7. Click the red arrow next to the column name and select **Test Probabilities**
8. Under **Test Probabilities** input the hypothesized probabilities for each category



Level	Estim Prob	Hypoth Prob
Sunday	0.14286	0.1429
Monday	0.19048	0.1429
Tuesday	0.09524	0.1429
Wednesday	0.14286	0.1429
Thursday	0.19048	0.1429
Friday	0.14286	0.1429
Saturday	0.09524	0.1429

Click then Enter Hypothesized Probabilities.

Choose rescaling method to sum probabilities to 1.
○ Fix omitted at estimated values, rescale hypothesis
● Fix hypothesized values, rescale omitted

9. Click **Done**

Copyright 2025 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s). Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

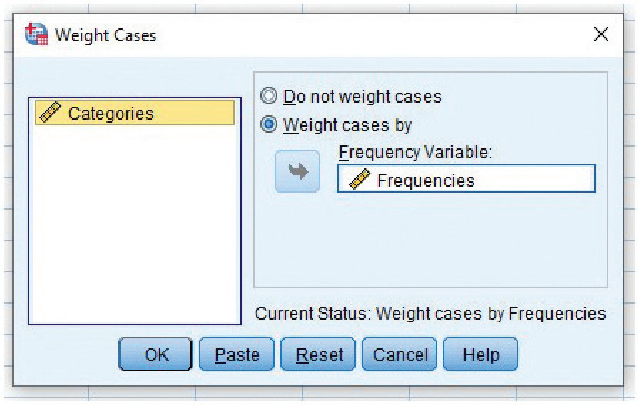
Note: The test statistic and *P*-value for the chi-squared test will appear in the row called **Pearson**.

Minitab

1. Enter each category name into the first column
2. Enter the count for each category into the second column
3. If the test involves unequal proportions, enter the hypothesized proportion for each category in the third column
4. Click **Stat** then click **Tables** then click **Chi-Square Goodness-of-Fit Test (One Variable)…**
5. Click in the box next to **Observed counts:**
6. Double-click the column name where the frequency counts are stored
7. Click in the box next to **Category names (optional):**
8. Double-click the column name where the category names are stored
9. For the test with **equal proportions**, click the radio button next to Equal proportions, then click **OK**
10. For the test with unequal proportions, click the radio button next to **Specific Proportions**. Click in the box next to **Specific Proportions** and double-click the column name where the hypothesized proportions are stored. Click **OK**.

SPSS

1. Number the categories from 1 to *k*. Input the numbers 1 to *k* into one column.
2. (Optional) In Variable View, under Values, define the category names represented by the numbers 1 to *k*
3. Input the frequency counts for each category into the second column
4. Click **Data** then click **Weight Cases…**
5. Click the radio button for **Weight cases by**. Click the name for the column containing the frequency counts and click the arrow to move this variable to the **Frequency Variable**: box.
6. Click **OK**



○ Do not weight cases
● Weight cases by
Frequency Variable: Frequencies
Current Status: Weight cases by Frequencies

Source: SPSS

7. Click **Analyze** then click **Nonparametric Tests** then click **Legacy Dialogs** then click **Chi-square…**
8. Click the name of the column containing the numbers 1 through *k* and click the arrow to move this variable to the **Test Variable** List box
9. For the test with equal proportions, click the radio button for **All categories equal** then click **OK**
10. For the test with unequal proportions, click the radio button for **Values:** Enter the hypothesized proportion for the first category into the box next to **Values:** then click **Add** to move it to the box next to **Add**. Repeat for the rest of the *k* categories. Click **OK**.

Excel

Excel does not have the functionality to produce the X^2 goodness-of-fit test automatically. Use Excel to find the *P*-value for this test after finding the value of the test statistic by using the following steps.

1. Click on an empty cell
2. Click **Formulas**
3. Click Insert Function
4. Select **Statistical** from the drop-down box for category
5. Select **CHISQ.DIST** and press **OK**
6. In the box next to **X**, type the value of the test statistic
7. In the box next to **Deg _freedom** type the value for the degrees of freedom
8. In the box next to **Cumulative** type **TRUE**
9. Click **OK**

Note: This outputs the value for $P(X \le x)$. To obtain the right tailed probability $P(X \ge x)$ click in an empty cell. Type = **1** − then click on the cell containing the cumulative probability and press **ENTER**.

TI-83/84

1. Enter the observed cell counts into **L1** and the expected cell counts into **L2** (in order to access lists press the **STAT** key, highlight the option called **Edit…** then press **ENTER**)
2. Press **STAT**
3. Highlight **TESTS**
4. Highlight **X^2GOF-Test** and press **ENTER**
5. Next to Observed enter **L1**
6. Next to Expected enter **L2**
7. Next to df enter the appropriate df (this will be the number of categories − 1)
8. Highlight **Calculate** and press **ENTER**

TI-Nspire

1. Enter the observed data into a data list (In order to access data lists select the spreadsheet option and press **enter**)

   Note: Be sure to title the lists by selecting the top row of the column and typing a title.

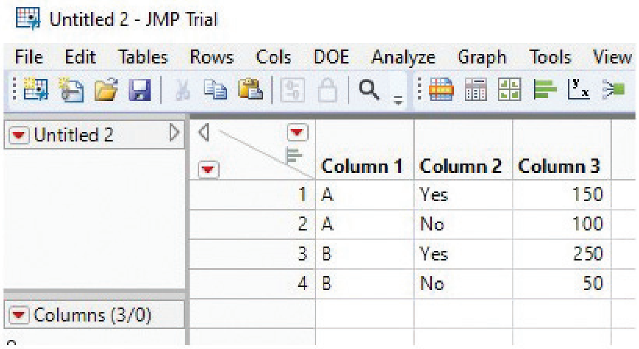
2. Enter the expected data into a data list
3. Press **menu** and select **4:Statistics** then **4:Stat Tests** then **7:** χ^2 **GOF…** and press **enter**
4. For **Observed List** select the title of the list that contains the observed data from the drop-down menu
5. For **Expected List** select the title of the list that contains the expected data from the drop-down menu
6. For **Deg of Freedom** input the degrees of freedom for this test
7. Press **OK**

Copyright 2025 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s). Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

X^2 Tests for Independence and Homogeneity

JMP

1. Input the data table into the JMP data table
2. Click **Analyze** then select **Fit Y by X**
3. Click and drag the first column name from the box under **Select Columns** to the box next to **Y, Response**



	Column 1	Column 2	Column 3
1	A	Yes	150
2	A	No	100
3	B	Yes	250
4	B	No	50

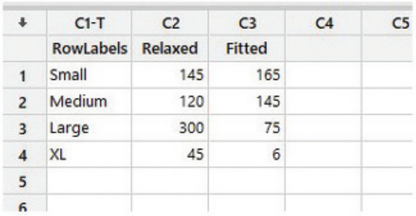
Source: JMP

4. Click and drag the second column name from the box under **Select Columns** to the box next to **X, Factor**
5. Click and drag the third column name from the box under **Select Columns** to the box next to **Freq**
6. Click **OK**

Note: The test statistic and *P*-value for the chi-squared test will appear in the row called **Pearson.**

Minitab

1. Input the data table into MINITAB



↓	C1-T	C2	C3	C4	C5
	RowLabels	Relaxed	Fitted		
1	Small	145	165		
2	Medium	120	145		
3	Large	300	75		
4	XL	45	6		
5					
6					

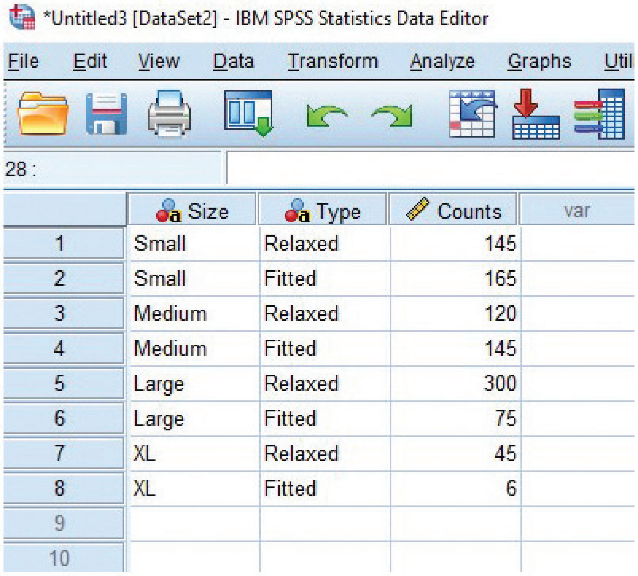
Source: Minitab

2. Click **Stat** then **Tables** then **Chi-Square Test for Association...**
3. Select **Summarized data in a two-way table** from the drop-down menu
4. Click in the box under **Columns containing the table:**
5. Double-click each of the column names containing the category counts
6. Click in the box next to **Rows:**
7. Double-click the column name where the row labels are stored
8. Click in the box next to **Columns:**
9. Type the name of the categorical variable that defines the table columns
10. Click **OK**

Note: This output returns expected cell counts as well as the chi-square test statistic and *P*-value.

SPSS

1. Enter the row variable data into one column
2. Enter the column variable data into a second column


*Untitled3 [DataSet2] - IBM SPSS Statistics Data Editor
File Edit View Data Transform Analyze Graphs Util
28 :

	Size	Type	Counts	var
1	Small	Relaxed	145	
2	Small	Fitted	165	
3	Medium	Relaxed	120	
4	Medium	Fitted	145	
5	Large	Relaxed	300	
6	Large	Fitted	75	
7	XL	Relaxed	45	
8	XL	Fitted	6	
9				
10				

Source: SPSS

3. Enter the frequency counts for each combination of categories into a third column.
4. Click **Data** then click **Weight Cases...**
5. Click the radio button for **Weight Cases by**. Click the name for the column containing the frequency counts and click the arrow to move this variable to the **Frequency Variable:** box.
6. Click **OK**

Note: steps 1 through 6 take a two-way contingency table and enter it into SPSS

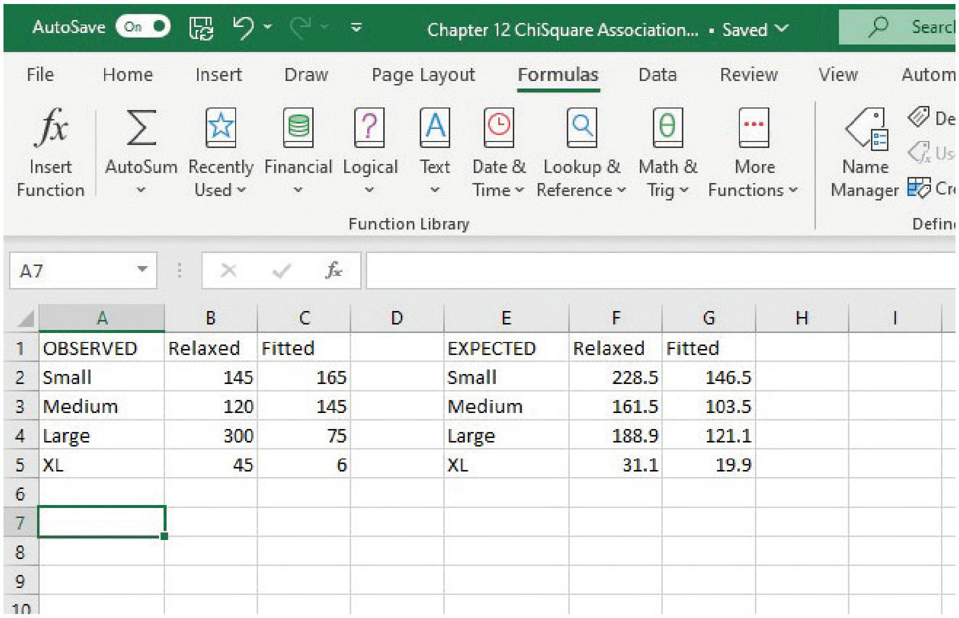
7. Click **Analyze** then click **Descriptive Statistics** then click **Crosstabs...**
8. Select the name of the row variable and press the arrow to move the variable to the box under **Row(s):**
9. Select the name of the column variable and press the arrow to move the variable to the box under **Column(s):**
10. Click **Statistics...**
11. Check the box next to **Chi-Square**
12. Click **Continue**
13. Click **Cells...**
14. Check the box next to **Expected**
15. Click **Continue**
16. Click **OK**

Note: The *P*-value for this test can be found in the **Chi-Square Tests** table in the Pearson Chi-Square row.

Excel

1. Input the observed contingency table
2. Input the expected table
3. Click on an empty cell
4. Click **Formulas**
5. Click **Insert Function**
6. Select **Statistical** from the drop-down box for category
7. Select **CHISQ.TEST** and press **OK**
8. Click in the box next to **Actual_range** and select the data values from the actual table (do NOT select column or row labels or totals)

Excel Worksheet Content After Step 4 Above

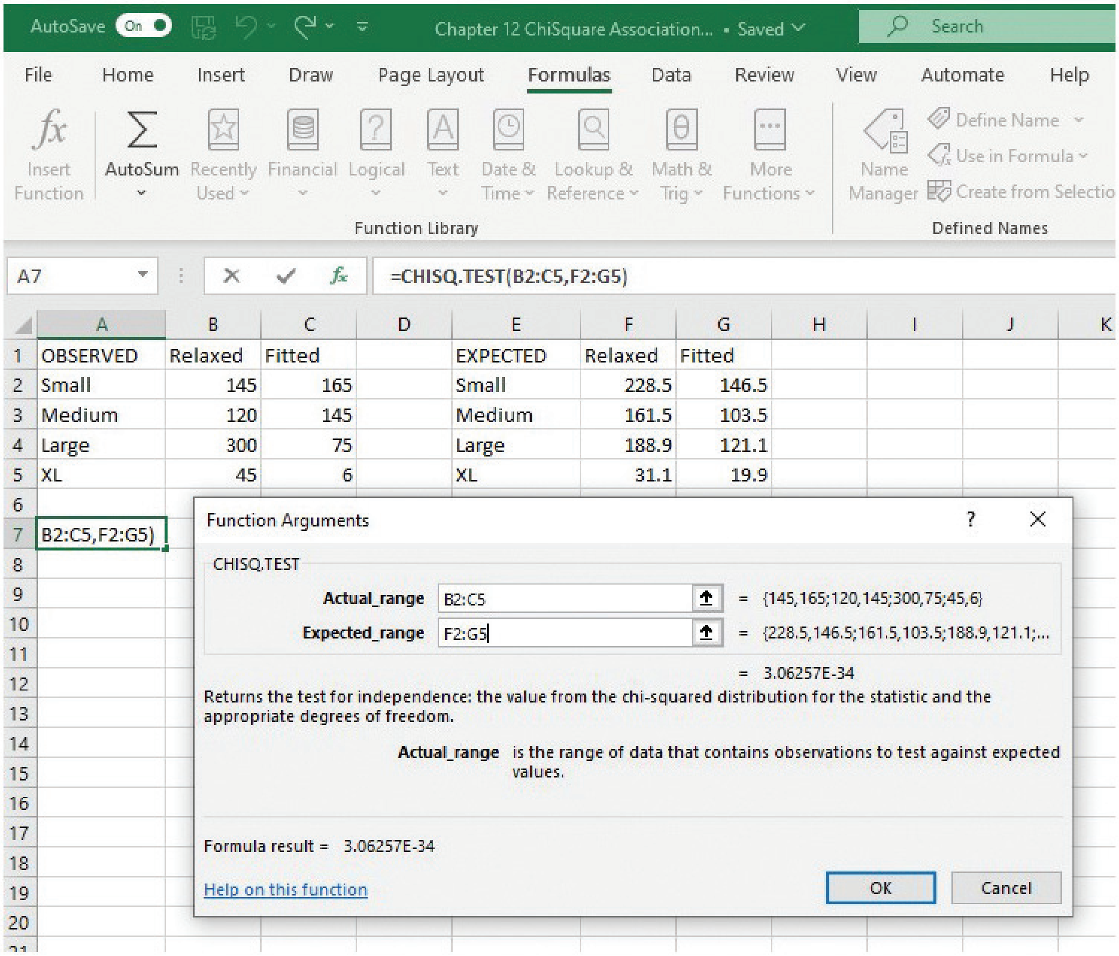

AutoSave On
Chapter 12 ChiSquare Association... • Saved
Search
File
Home
Insert
Draw
Page Layout
Formulas
Data
Review
View
Autom
Insert Function
AutoSum
Recently Used
Financial
Logical
Text
Date & Time
Lookup & Reference
Math & Trig
More Functions
Function Library
Name Manager
De
Us
Cr
Defin

A7

	A	B	C	D	E	F	G
1	OBSERVED	Relaxed	Fitted		EXPECTED	Relaxed	Fitted
2	Small	145	165		Small	228.5	146.5
3	Medium	120	145		Medium	161.5	103.5
4	Large	300	75		Large	188.9	121.1
5	XL	45	6		XL	31.1	19.9

Source: Microsoft Corporation

Excel Worksheet Content After Step 10 on Page 725


AutoSave On
Chapter 12 ChiSquare Association... • Saved
Search
File
Home
Insert
Draw
Page Layout
Formulas
Data
Review
View
Automate
Help
Insert Function
AutoSum
Recently Used
Financial
Logical
Text
Date & Time
Lookup & Reference
Math & Trig
More Functions
Function Library
Name Manager
Define Name
Use in Formula
Create from Selectio
Defined Names

A7 =CHISQ.TEST(B2:C5,F2:G5)

	A	B	C	D	E	F	G
1	OBSERVED	Relaxed	Fitted		EXPECTED	Relaxed	Fitted
2	Small	145	165		Small	228.5	146.5
3	Medium	120	145		Medium	161.5	103.5
4	Large	300	75		Large	188.9	121.1
5	XL	45	6		XL	31.1	19.9
7	B2:C5,F2:G5)						

Function Arguments

CHISQ.TEST

Actual_range B2:C5 = {145,165;120,145;300,75;45,6}

Expected_range F2:G5 = {228.5,146.5;161.5,103.5;188.9,121.1;...

= 3.06257E-34

Returns the test for independence: the value from the chi-squared distribution for the statistic and the appropriate degrees of freedom.

Actual_range is the range of data that contains observations to test against expected values.

Formula result = 3.06257E-34

Help on this function

OK Cancel

Source: Microsoft Corporation

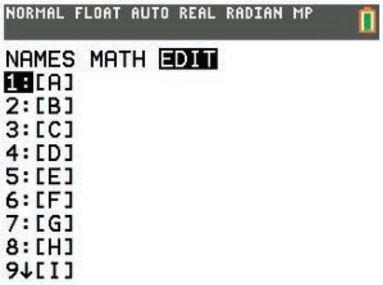
Copyright 2025 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s). Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

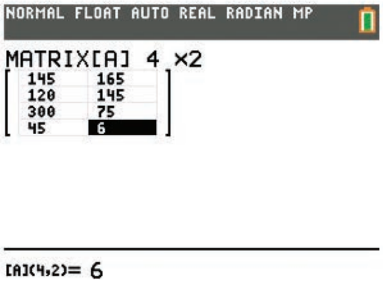
9. Click in the box next to **Expected_range** and select the data values from the expected table (do NOT select column or row labels or totals)
10. Click **OK**

Note: This outputs the value for $P(X \geq x)$.

TI-83/84

1. Input the observed contingency table into matrix **A** (To access and edit matrices, press **2nd** then $\mathbf{x^{-1}}$, then highlight **EDIT**. Then highlight **[A]** and press **ENTER**. Type the value for the number of rows and press **ENTER**; type the value for the number of columns and press **ENTER**. Type the data values into the matrix.)



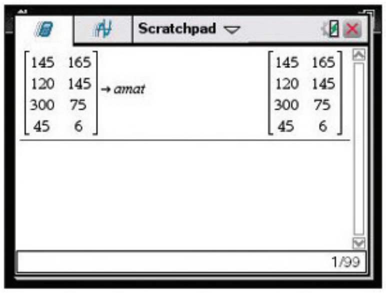


```
MATRIX[A] 4 ×2
145   165
120   145
300   75
45    6

[A](4,2)= 6
```

2. Input the expected contingency table into matrix **B** (Follow same instructions as for matrix **A** except highlight [**B**].)
3. Press **STAT**
4. Highlight **TESTS**
5. Highlight χ^2**-Test…** and press **ENTER**
6. Highlight **Calculate** and press **ENTER**

TI-Nspire

1. Enter the Calculator Scratchpad
2. Press the **menu** key then select **7:Matrix & Vector** then select **1:Create** then select **1:Matrix…** and press **enter**
3. Next to **Number of rows** enter the number of rows in the contingency table (do not include title rows or total rows)
4. Next to **Number of columns** enter the number of columns in the contingency table (do not include title columns or total columns)
5. Press **OK**
6. Input the values into the matrix (pressing tab after entering each value and press enter)
7. Press **ctrl** then press **var**
8. Type in **amat** and press **enter**



$$\begin{bmatrix} 145 & 165 \\ 120 & 145 \\ 300 & 75 \\ 45 & 6 \end{bmatrix} \rightarrow amat \qquad \begin{bmatrix} 145 & 165 \\ 120 & 145 \\ 300 & 75 \\ 45 & 6 \end{bmatrix}$$

Source: TI-Nspire

9. Press the **menu** key then select **6:Statistics** then **7:Stat tests** then **8:** χ^2 **2-way Test…**
10. For **Observed Matrix**, select amat from the drop-down list
11. Press **OK**

Copyright 2025 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s). Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

Chapter 13

Simple Linear Regression and Correlation: Inferential Methods

Monkey Business Images/Shutterstock.com

Regression and correlation were introduced in Chapter 5 as methods for describing and summarizing bivariate numerical data consisting of (x, y) pairs. For example, data on

y = College grade point average (GPA)

and

x = Language processing score

for a sample of older college students (age 50 to 79) were used in an investigation of predictors of academic success for older students **("Age Is No Barrier: Predictors of Academic Success in Older Learners,"** ***Nature Partner Journals*****, nature.com, 2017)**. The data were used to construct a scatterplot. The scatterplot showed a linear pattern. The sample correlation coefficient was $r = 0.323$, and the equation of the least-squares line had a positive slope, indicating that older students with higher language processing scores tended to have higher GPAs.

Could the pattern observed in the scatterplot be plausibly explained by chance, or does the sample provide convincing evidence that there really is a linear relationship between these two variables for older college students? If there is evidence of a meaningful relationship between these two variables, the regression line could be used as the basis for predicting the GPA for an older student with a specified language processing score.

In this chapter, inferential methods for bivariate numerical data are introduced, including a hypotheses test to determine if there is a useful linear relationship in a population of (x, y) pairs.

Learning Objectives

After completing this chapter, you will be able to:

- Interpret the parameters of the simple linear regression model in context.
- Assess whether the assumptions of the simple linear regression model are reasonable using scatterplots, residual plots, and normal probability plots.
- Construct and interpret a confidence interval for the slope of the population regression line.
- Test hypotheses about the slope of the population regression line and interpret the results in context.

Copyright 2025 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s). Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

Section 13.1 Simple Linear Regression Model

A **deterministic relationship** is one in which the value of y is completely determined by the value of an independent variable x. A deterministic relationship can be described using traditional mathematical notation $y = f(x)$, where $f(x)$ is a specified function of x. For example,

$$y = f(x) = 10 + 2x$$

or

$$y = f(x) = 4 - (10)^{2x}$$

However, in many situations, the variables of interest are not deterministically related. For example, the value of y = first-year college grade point average is not an exact function of x = number of hours spent studying per week, and y = crop yield is determined partly by factors other than x = amount of fertilizer used.

What Do You Think?

Consider the following two variables:

x = height of an elementary school student
y = age

Is the relationship between x and y a deterministic relationship?

A relationship between two variables x and y that are not deterministically related is described by a **probabilistic model**. The general form of an **additive probabilistic model** allows y to be larger or smaller than $f(x)$ by a random amount e. The **model equation** is of the form

$$y = \text{deterministic function of } x + \text{random deviation} = f(x) + e$$

For example, consider the probabilistic model

$$y = \underbrace{50 - 10x + x^2}_{f(x)} + e$$

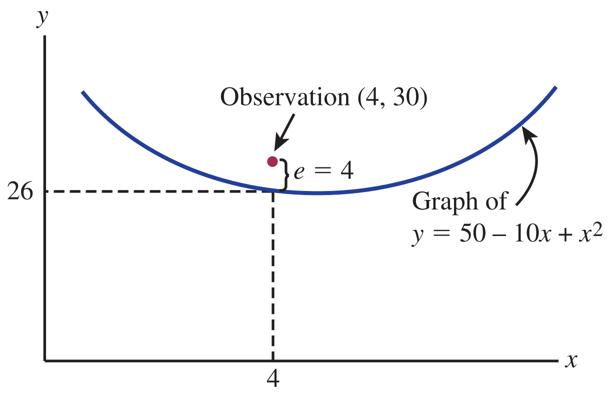
The graph of the function $y = 50 - 10x + x^2$ is shown as the curve in Figure 13.1. The observed point (4, 30) is also shown in the figure. Because

$$f(4) = 50 - 10(4) + 4^2 = 50 - 40 + 16 = 26$$

for the point (4, 30), you can write $y = f(x) + e$, where $e = 4$. The point (4, 30) falls 4 above the graph of the function $y = 50 - 10x + x^2$.

Thinking geometrically, if $e > 0$, the corresponding point will lie above the graph of $y = f(x)$. If $e < 0$ the corresponding point will fall below the graph. If $f(x)$ is a function used in a probabilistic model relating y to x and if observations on y are made for various values of x, the resulting (x, y) points will be distributed around the graph of $f(x)$, some falling above it and some falling below it.

Figure 13.1
A deviation from the deterministic part of a probabilistic model.



Simple Linear Regression

The simple linear regression model is a special case of the general probabilistic model in which the deterministic function $f(x)$ is linear (so its graph is a straight line).

Definitions

Simple linear regression model: A probabilistic model in which the deterministic part is a line with vertical or y intercept α and slope β. This line is called the **population regression line**.

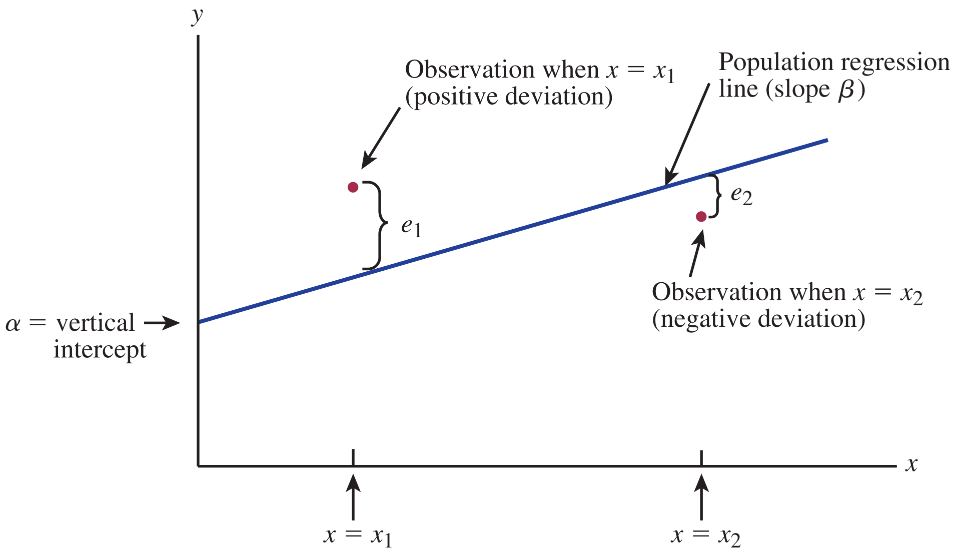
When a value of the independent variable x is fixed and an observation on the dependent variable y is made,

$$y = \alpha + \beta x + e$$

Without the random deviation e, all observed (x, y) points would fall exactly on the population regression line. The inclusion of e in the model equation recognizes that points will deviate from the line by a random amount.

Figure 13.2 shows two observations in relation to a population regression line.

Figure 13.2
Two observations and deviations from a population regression line.



y
Observation when $x = x_1$ (positive deviation)
Population regression line (slope β)
e_1
e_2
Observation when $x = x_2$ (negative deviation)
α = vertical intercept
x
$x = x_1$
$x = x_2$

Before an observation of y for any particular value of x is made, there is uncertainty about the value of e. It could be negative, positive, or even 0. Also, it might be quite large in magnitude (a point far from the population regression line) or quite small (a point very close to the line). In this chapter, the following assumptions about the distribution of e in repeated sampling at any particular x value are made.

Basic Assumptions of the Simple Linear Regression Model

1. The distribution of e at any particular x value has mean value 0. That is, $\mu_e = 0$.
2. The standard deviation of e (which describes the variability in its distribution) is the same for any particular value of x. This standard deviation is denoted by σ.
3. The distribution of e at any particular x value is normal.
4. The random deviations $e_1, e_2, \ldots, e_n$ associated with different observations are independent of one another.

Copyright 2025 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s). Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

These assumptions about the e term in the simple linear regression model also imply that there is variability in the y values observed at any particular value of x. Consider y when x has some fixed value x^*, so that

$$y = \alpha + \beta x^* + e$$

Because α and β are fixed numbers, $\alpha + \beta x^*$ is also a fixed number. The sum of a fixed number and a normally distributed variable (e) is also a normally distributed variable (the bell-shaped curve is simply shifted). This means that y itself has a normal distribution. Furthermore, $\mu_e = 0$ implies that the mean value of y is $\alpha + \beta x^*$, the height of the population regression line above the value x^*. Finally, because there is no variability in the fixed number $\alpha + \beta x^*$, the standard deviation of y is the same as the standard deviation of e. These properties are summarized in the following box.

At any fixed x value, y has a normal distribution, with

$$\begin{pmatrix}\text{mean } y \text{ value}\\ \text{for fixed } x\end{pmatrix} = \begin{pmatrix}\text{height of the population}\\ \text{regression line above } x\end{pmatrix} = \alpha + \beta x$$

and

$$(\text{standard deviation of } y \text{ for a fixed } x) = \sigma$$

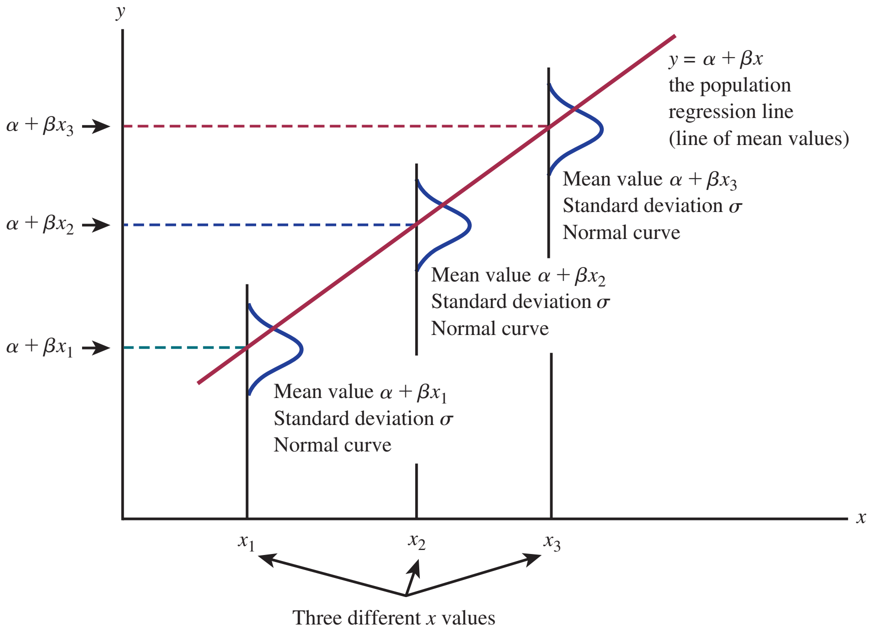
The slope β of the population regression line is the *average* change in y associated with a 1-unit increase in x.

The y intercept α is the height of the population line when $x = 0$.

The value of σ determines the extent to which (x, y) observations deviate from the population line. When σ is small, most observations will be quite close to the line, but when σ is large, there are likely to be some large deviations.

The key features of the model are illustrated in Figures 13.3 and 13.4. Notice that the three normal curves in Figure 13.3 have the same standard deviation (equal variability). This is a consequence of σ_e being the same at any value of x, which implies that the variability in the y values at a particular x does not depend on the value of x.

Figure 13.3
Illustration of the simple linear regression model.


y
$\alpha + \beta x_3$
$\alpha + \beta x_2$
$\alpha + \beta x_1$
$y = \alpha + \beta x$
the population
regression line
(line of mean values)
Mean value $\alpha + \beta x_3$
Standard deviation σ
Normal curve
Mean value $\alpha + \beta x_2$
Standard deviation σ
Normal curve
Mean value $\alpha + \beta x_1$
Standard deviation σ
Normal curve
x
x_1
x_2
x_3
Three different x values

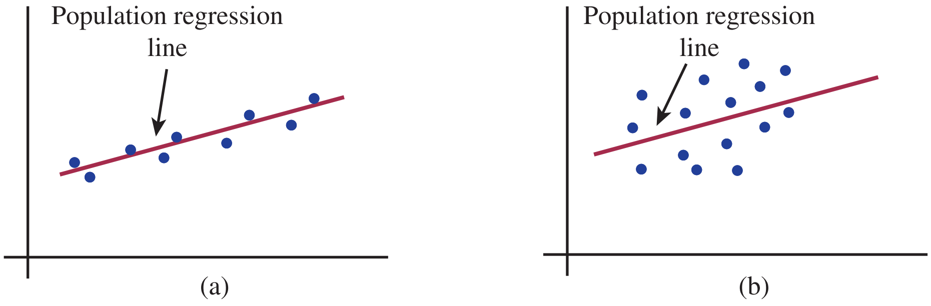


Figure 13.4
A simple linear regression model:
(a) small σ;
(b) large σ.

Example 13.1 Text Neck—Keep Your Head Up!

Understand the context ➤

The article **"Text Neck Is Becoming an Epidemic and Could Wreck Your Spine" (*Washington Post*, November 20, 2014)** reports that the human head weighs about 12 pounds but that as the neck bends forward and down, pressure on the cervical spine increases as the angle of the neck increases. This is a concern because so many people spend a lot of time looking down at their phones and other electronic devices and this could lead to additional wear-and-tear to the spine.

A linear regression model can be used to describe the relationship between the variables

x = Forward angle of the neck (0 corresponds to head upright)

and

y = Pressure on the cervical spine (in pounds)

Suppose that the linear regression model has $\alpha = 14$, $\beta = 0.8$, and $\sigma_e = 2.1$ (these values are consistent with values given in the article). The population regression line is shown in Figure 13.5.

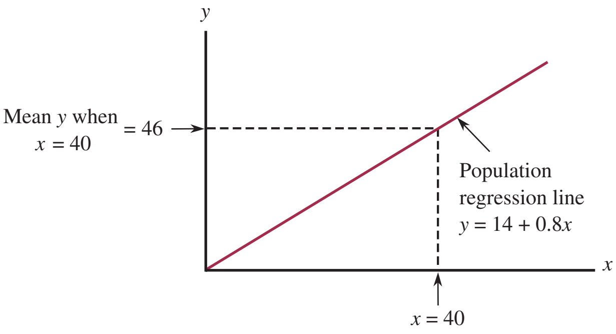


Figure 13.5
The population regression line for Example 13.1

If the distribution of the random errors at any fixed weight (x value) is normal, then the variable y = Pressure on cervical spine is normally distributed with

$$\mu_y = 14 + 0.8x$$
$$\sigma_y = \sigma_e = 2.1$$

For example, when $x = 40$ (corresponding to people with a forward head angle of 40 degrees), pressure on the cervical spine has mean value

$$\mu_y = 14 + 0.8(40) = 46 \text{ pounds}$$

Because the standard deviation of y is $\sigma_y = 2.1$, the interval $46 \pm 2(2.1) = (41.8, 50.2)$ includes y values that are within 2 standard deviations of the mean value for y when $x = 40$. Roughly 95% of people with a forward head angle of 40 degrees will have a cervical spine pressure in this range. The slope $\beta = 0.8$ can be interpreted as the *mean* change

Copyright 2025 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s). Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

in pressure on the cervical spine associated with each additional degree of forward head angle. Because increased pressure on the neck is not a good thing, the article recommends that people make an effort to keep their heads as upright as possible when looking at their phones and avoid spending hours hunched over electronic devices.

More insight into model properties can be gained by thinking of the population of all (x, y) pairs as consisting of many smaller populations. Each one of these smaller populations contains pairs for which x has a fixed value. For example, suppose that in a large population of college students the relationship between the variables

x = Grade point average in major courses

and

y = Starting salary after graduation

can be described by the simple linear regression model. Then there is the population of all pairs with $x = 3.20$ (corresponding to all students with a grade point average of 3.20 in major courses), the population of all pairs having $x = 2.75$, and so on. The model assumes that for each such population, y is normally distributed with the same standard deviation, and the *mean y value* (rather than y itself) is a linear function of x.

In practice, the initial judgment of whether the simple linear regression model is appropriate is based on how the data were collected and on a scatterplot of the data. The sample observations should be independent of one another. In addition, the scatterplot should show a linear rather than a curved pattern, and the vertical variability of points should be relatively homogeneous throughout the range of x values. Figure 13.6 shows plots with three different patterns, but only the first is consistent with the simple liner regression model assumptions.

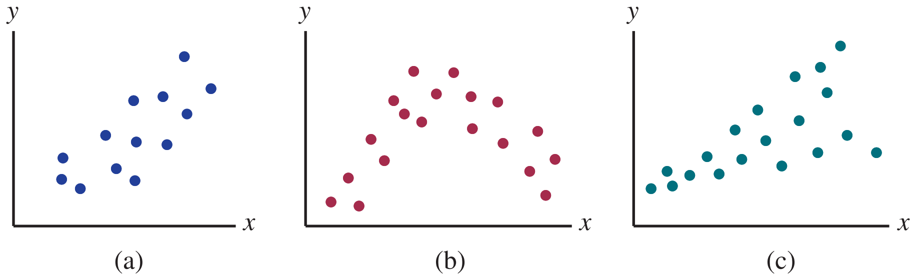


Figure 13.6
Some commonly encountered patterns in scatterplots:
(a) consistent with the simple linear regression model;
(b) suggests a nonlinear probabilistic model;
(c) suggests that variability in y changes with x.

Estimating the Slope and Intercept of the Population Regression Line

For the remainder of this chapter, the basic assumptions of the simple linear regression model will be presumed to be reasonable. The values of α and β (the y intercept and the slope of the population regression line) will almost never be known to an investigator. Instead, these values must be estimated from the sample data.

The letters a and b are used to denote the point estimates of α and β, respectively. These estimates are based on the method of least squares introduced in Chapter 5. The sum of squared vertical deviations is smaller for the least-squares line than for any other line.

The point estimates of β, the slope, and α, the y intercept of the population regression line, are the slope and y intercept of the least-squares line:

$$b = \text{point estimate of } \beta = \frac{S_{xy}}{S_{xx}}$$

$$a = \text{point estimate of } \alpha = \bar{y} - b\bar{x}$$

where

$$S_{xy} = \Sigma xy - \frac{(\Sigma x)(\Sigma y)}{n} \quad \text{and} \quad S_{xx} = \Sigma x^2 - \frac{(\Sigma x)^2}{n}$$

The estimated population regression line is the least-squares line

$$\hat{y} = a + bx$$

Suppose x^* denotes a specified value of the predictor variable x. Then $a + bx^*$ has two different interpretations:

1. It is a point estimate of the mean y value when $x = x^*$.
2. It is a point prediction of an individual y value to be observed when $x = x^*$.

Example 13.2 Roller Coaster Speed and Height

Understand the context ➤

● It makes sense to think that the top speed attained by a roller coaster may be related to how tall the roller coaster is at its highest point. Data from the Roller Coaster Data Base can be used to investigate this relationship. The following data on

y = Top speed (in miles per hour)

and

x = Maximum height (in feet)

are for a random sample of 10 roller coasters in the United States.

Consider the data ➤

Top Speed (mph)	Maximum Height (feet)
65	150
55	70
62	125
60	145
60	79
65	159
55	83
57	105
70	188
75	200

A scatterplot of the data is given in Figure 13.7. The scatterplot shows a linear pattern and the spread in the y values appears to be similar across the range of the x values. This supports the appropriateness of the simple linear regression model.

● **Data set available online**

Copyright 2025 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s). Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

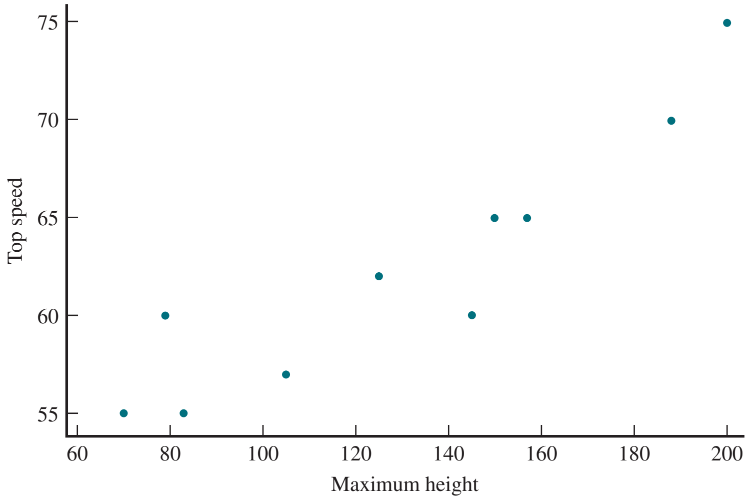


Figure 13.7
Scatterplot of top speed versus maximum height for Example 13.2

Do the work ➤ For these data, the equation of the estimated regression line was found using statistical software, resulting in

$$\hat{y} = a + bx = 45.25 + 0.13x$$

An estimate of the mean top speed for roller coasters that have a maximum height of 100 feet is found by substituting $x = 100$ into the estimated regression equation:

$$\begin{aligned}\text{estimated mean } y \text{ for roller coasters} \\ \text{with maximum height of 100 feet} &= a + bx \\ &= 45.25 + 0.13(100) \\ &= 58.25 \text{ miles per hour}\end{aligned}$$

Similarly, you would predict the top speed of an individual roller coaster with maximum height of 100 feet to be

$$\begin{aligned}\hat{y} &= \text{predicted } y \text{ value when } x = 100 \\ &= a + bx \\ &= 45.25 + 0.13(100) \\ &= 58.25 \text{ miles per hour}\end{aligned}$$

Interpret the results ➤ The estimate of the mean top speed and the prediction of an individual top speed are identical because the same *x* value was used in each calculation. However, their interpretations are different. One is the prediction of the top speed of a single roller coaster that has a maximum height of 100 feet, while the other is an estimate of the mean top speed of all roller coasters with a maximum height of 100 feet.

In Example 13.2, the *x* values in the sample ranged from 70 to 200. An estimate or prediction should not be attempted for any *x* value much outside this range. Without sample data for such values, there is no evidence that the observed linear relationship continues outside the range from 70 to 200. For this reason, it is not a good idea to make predictions outside this range. Statisticians refer to this potential pitfall as the **danger of extrapolation**.

Estimating σ^2 and σ

The value of σ describes the extent to which observed points (x, y) tend to fall close to or far away from the population regression line. A point estimate of σ is based on

$$SSResid = \sum(y - \hat{y})^2$$

where $\hat{y}_1 = a + bx_1, \ldots, \hat{y}_n = a + bx_n$ are the fitted or predicted *y* values and the residuals are $y_1 - \hat{y}_1, \ldots, y_n - \hat{y}_n$. SSResid is a measure of the extent to which the sample data spread out about the estimated regression line.

The statistic for estimating the variance σ^2 is

$$s_e^2 = \frac{\text{SSResid}}{n-2}$$

where

$$\text{SSResid} = \Sigma(y-\hat{y})^2 = \Sigma y^2 - a\Sigma y - b\Sigma xy$$

The subscript e in s_e^2 reminds you that you are estimating the variance of the "errors" or residuals.

The estimate of σ is the **estimated standard deviation**

$$s_e = \sqrt{s_e^2}$$

The number of degrees of freedom associated with estimating σ^2 or σ in simple linear regression is $n-2$.

The estimates and number of degrees of freedom here have analogs in previous work involving a single sample $x_1, x_2, \ldots, x_n$. The sample variance s^2 had a numerator of $\Sigma(x-\bar{x})^2$, a sum of squared deviations (residuals), and denominator $n-1$, the number of degrees of freedom associated with s^2 and s. The use of $\bar{x}$ as an estimate of μ in the formula for s^2 reduces the number of degrees of freedom by 1, from n to $n-1$. In simple linear regression, estimation of α and β results in a loss of 2 degrees of freedom, leaving $n-2$ as the number of degrees of freedom for SSResid, s_e^2, and s_e.

The coefficient of determination was defined previously (see Chapter 5) as

$$r^2 = 1 - \frac{\text{SSResid}}{\text{SSTo}}$$

where

$$\text{SSTo} = \Sigma(y-\bar{y})^2 = \Sigma y^2 - \frac{(\Sigma y)^2}{n} = S_{yy}$$

The value of r^2 is interpreted as the proportion of observed variability in y that can be explained by (or attributed to) the model relationship.

The estimate s_e also gives another assessment of model performance. Roughly speaking, the value of σ represents the magnitude of a typical deviation of a point (x, y) in the population from the population regression line. Similarly, s_e is the approximate magnitude of a typical sample deviation (residual) from the least-squares line. The smaller the value of s_e, the closer the points in the sample tend to fall to the line and the better you would expect the line to do in predicting y from x.

Example 13.3 Predicting Election Outcomes

Understand the context ➤

● The authors of the paper **"Inferences of Competence from Faces Predict Election Outcomes" (*Science* [2005]: 1623–1626)** found that they could successfully predict the outcome of a U.S. congressional election substantially more than half the time based on the facial appearance of the candidates. In the study described in the paper, participants were shown photos of two candidates for a U.S. Senate or House of Representatives election. Each participant was asked to look at the photos and then indicate which candidate they thought was more competent.

The two candidates were labeled A and B. If a participant recognized either candidate, data from that participant were not used in the analysis. The proportion of participants who chose candidate A as the more competent was calculated. After the election, the difference in votes (candidate A − candidate B) expressed as a proportion of the total votes cast in

● **Data set available online**

the election was also calculated. This difference falls between +1 and −1. It is 0 for an election where both candidates receive the same number of votes, positive for an election where candidate A received more votes than candidate B (with +1 indicating that candidate A received all of the votes), and negative for an election where candidate A received fewer votes than candidate B.

This process was carried out for a large number of congressional races. A subset of the resulting data (approximate values read from a graph that appears in the paper) is given in the accompanying table, which also includes the predicted values and residuals for the least-squares line for these data.

Consider the data ➤

Competent Proportion	Difference in Vote Proportion	Predicted *y* Value	Residual
0.20	−0.70	−0.389	−0.311
0.23	−0.40	−0.347	−0.053
0.40	−0.35	−0.109	−0.241
0.35	0.18	−0.179	0.359
0.40	0.38	−0.109	0.489
0.45	−0.10	−0.040	−0.060
0.50	0.20	0.030	0.170
0.55	−0.30	0.100	−0.400
0.60	0.30	0.170	0.130
0.68	0.18	0.281	−0.101
0.70	0.50	0.309	0.191
0.76	0.22	0.393	−0.173

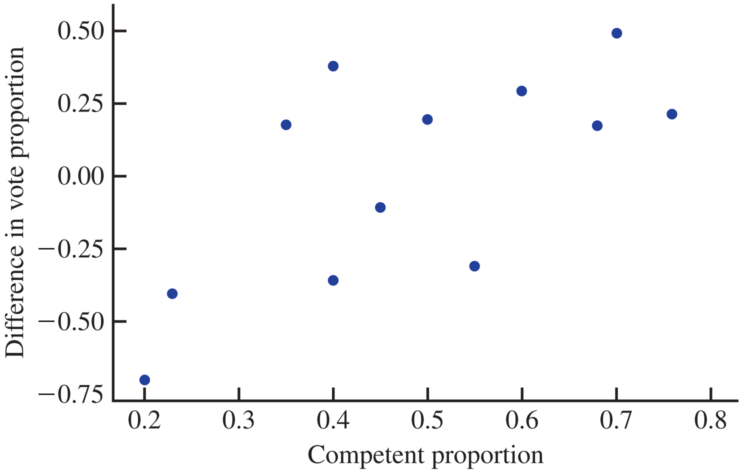
The scatterplot (Figure 13.8) suggests a positive linear relationship between

x = Proportion of participants who judged candidate A as the more competent

and

y = Difference in vote proportion.

Figure 13.8
Minitab scatterplot for Example 13.3.



What Do You Think?

Looking at the scatterplot in Figure 13.8, do you think that the value of s_e is large or small?

Copyright 2025 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s). Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

Do the work ➤ The summary statistics are

$$n = 12 \qquad \Sigma x = 5.82 \qquad \Sigma y = 0.11$$

$$\Sigma x^2 = 3.1804 \quad \Sigma xy = 0.5526 \quad \Sigma y^2 = 1.5101$$

from which you can calculate

$$b = 1.3957 \qquad a = -0.6678$$

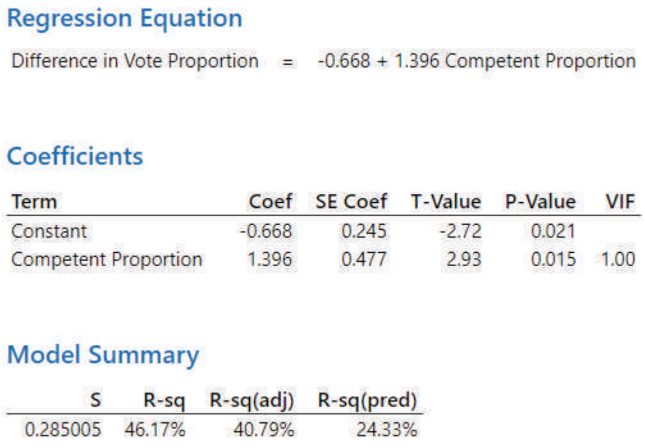
$$\text{SSResid} = 0.81228 \qquad \text{SSTo} = 1.50909$$

$$r^2 = 1 - \frac{\text{SSResid}}{\text{SSTo}} = 1 - \frac{0.81228}{1.50909} = 1 - 0.538 = 0.462$$

$$s_e^2 = \frac{\text{SSResid}}{n-2} = \frac{0.81228}{10} = 0.081$$

$$s_e = \sqrt{0.081} = 0.285$$

A statistical software package could also have been used to calculate the equation of the estimated regression line and the values of s_e and r^2. Minitab output is shown here.



```
Regression Equation

Difference in Vote Proportion  =  -0.668 + 1.396 Competent Proportion


Coefficients

Term                    Coef  SE Coef  T-Value  P-Value   VIF
Constant              -0.668    0.245    -2.72    0.021
Competent Proportion   1.396    0.477     2.93    0.015  1.00


Model Summary

        S    R-sq  R-sq(adj)  R-sq(pred)
 0.285005  46.17%     40.79%      24.33%
```

Interpret the results ➤ Approximately 46.2% of the observed variability in the difference in vote proportion *y* can be attributed to the linear relationship with proportion of participants who judged the candidate to be more competent based on facial appearance alone. The magnitude of a typical sample deviation from the least-squares line is about 0.285, which is not very large in comparison to the *y* values themselves. The model may be useful for estimation and prediction. In Section 13.2, you will see how a model utility test can be used to judge whether this is actually the case.

A key assumption of the simple linear regression model is that the random deviation *e* in the model equation is normally distributed. In Section 13.3, you will see how the residuals can be used to determine whether this is plausible.

Exercises 13.1–13.13

● **Data set available online**

13.1 Identify the following relationships as deterministic or probabilistic:

- **a.** The relationship between the length of the sides of a square and its perimeter.
- **b.** The relationship between the height and weight of an adult.
- **c.** The relationship between score on a placement exam and first-year college GPA.
- **d.** The relationship between tree height in centimeters and tree height in inches.

Copyright 2025 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s). Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

13.2 Let x be the size of a house (in square feet) and y be the amount of natural gas used (therms) during a specified period. Suppose that for a particular community, x and y are related according to the simple linear regression model with

β = slope of population regression line = 0.017
α = y intercept of population regression line = −5.0

Houses in this community range in size from 1000 to 3000 square feet.

a. What is the equation of the population regression line?
b. Graph the population regression line by first finding the point on the line corresponding to $x = 1000$ and then the point corresponding to $x = 2000$, and drawing a line through these points.

13.3 Consider the variables and population regression line given in the previous exercise.

a. What is the mean value of gas usage for houses with 2100 sq. ft. of space?
b. What is the average change in usage associated with a 1 sq. ft. increase in size?
c. What is the average change in usage associated with a 100 sq. ft. increase in size?
d. Should the model be used to predict mean usage for a 500 sq. ft. house? Why or why not? (Hint: See the discussion following Example 13.2.)

13.4 The paper **"Predicting Yolk Height, Yolk Width, Albumen Length, Eggshell Weight, Egg Shape Index, Eggshell Thickness, Egg Surface Area of Japanese Quails Using Various Egg Traits as Regressors" (*International Journal of Poultry Science* [2008]: 85–88)** suggests that the simple linear regression model is reasonable for describing the relationship between y = Eggshell thickness (in micrometers) and x = Egg length (mm) for quail eggs. Suppose that the population regression line is $y = 0.135 + 0.003x$ and that $\sigma = 0.005$. Then, for a fixed x value, y has a normal distribution with mean $0.135 + 0.003x$ and standard deviation 0.005.

a. What is the mean eggshell thickness for quail eggs that are 15 mm in length? For quail eggs that are 17 mm in length?
b. What is the probability that a quail egg with a length of 15 mm will have a shell thickness that is greater than 0.18 μm?
c. Approximately what proportion of quail eggs of length 14 mm have a shell thickness of greater than 0.175? Less than 0.178? (Hint: The distribution of y at a fixed x is approximately normal.)

13.5 A sample of small cars was selected, and the values of x = horsepower and y = fuel efficiency (mpg) were determined for each car. Fitting the simple linear regression model gave the estimated regression equation $\hat{y} = 44.0 - 0.150x$.

a. Interpret the value of $b = -0.150$.
b. Substituting $x = 100$ gives $\hat{y} = 29.0$. Give two different interpretations of this number.
c. What happens if the efficiency for a car with a 300-horsepower engine is predicted? Why does this occur?

13.6 Use the information given in the previous exercise to answer the following questions.

a. The value of r^2 was found to be 0.680. Interpret r^2 in the context of this problem.
b. The value of s_e was found to be 3.0. Interpret s_e in the context of this problem.

13.7 Suppose that a simple linear regression model is appropriate for describing the relationship between y = House price (in dollars) and x = House size (in square feet) for houses in a large city. The population regression line is $y = 23{,}000 + 47x$ and $\sigma = 5000$.

a. What is the average change in price associated with one extra square foot of space? With an additional 100 sq. ft. of space?
b. Approximately what proportion of 1800 sq. ft. homes would be priced over $110,000? Under $100,000?

13.8 **a.** Explain the difference between the line $y = \alpha + \beta x$ and the line $\hat{y} = a + bx$.
b. Explain the difference between β and b.
c. Let x^* denote a particular value of the independent variable. Explain the difference between $\alpha + \beta x^*$ and $a + bx^*$.
d. Explain the difference between σ and s_e.

13.9 ● The paper **"Depression, Body Mass Index, and Chronic Obstructive Pulmonary Disease—A Holistic Approach" (*International Journal of COPD* [2016]: 239–249)** gives data on

x = Change in body mass index (BMI in kg/m^2)

and

y = Change in a measure of depression

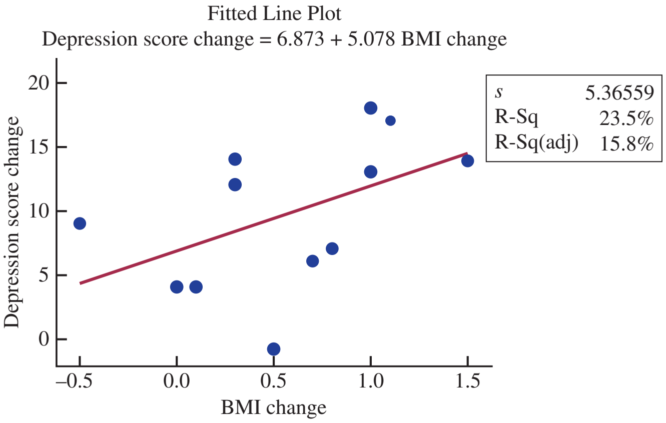
for patients suffering from depression who participated in a pulmonary rehabilitation program. The table below contains a subset of the data given in the paper and are approximate values read from a scatterplot in the paper.

BMI Change (kg/m²)	Depression Score Change
0.5	−1
−0.5	9
0.0	4
0.1	4
0.7	6
0.8	7
1.0	13
1.5	14

(continued)

BMI Change (kg/m²)	Depression Score Change
1.1	17
1.0	18
0.3	12
0.3	14

The accompanying computer output is from Minitab.



```
       S          R-sq
5.36559         23.48%

Coefficients
Term        Coef  SE Coef  T-Value  P-Value  VIF
Constant    6.87    2.26     3.04    0.012
BMI change  5.08    2.90     1.75    0.110 1.00

Regression Equation
Depression score change =
6.87 + 5.08 BMI change
```

a. What percentage of observed variation in depression score change can be explained by the simple linear regression model? (Hint: See Example 13.3.)
b. Give an estimate of σ and interpret this estimate.
c. Give an estimate of the average change in depression score change associated with a 1 kg/m^2 increase in BMI change.
d. Calculate a point estimate of the mean depression score change for a patient whose BMI change was 1.2 kg/m^2.

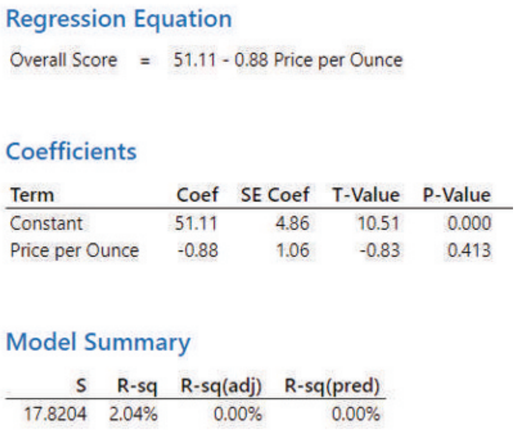
13.10 ● The ability to jump helps salamanders move quickly to avoid predators. The paper **"Jumping in Arboreal Salamanders: A Possible Tradeoff Between Takeoff Velocity and In-Air Posture" (*Zoology* [2020]: 125724)** investigated jumping behavior of 10 species of salamanders and reported the following data on bend angle (the bend [in degrees] in the salamander body, with a larger number reflecting more extreme bending) and the jump takeoff velocity (in meters per second).

Bend Angle	Takeoff Velocity
36.5	0.6
55.0	0.7
43.5	0.7
50.4	0.7
69.8	0.9
71.2	1.0
67.4	1.1
48.1	0.7
61.8	0.9
52.6	0.9

A simple linear regression model can be used to describe the relationship between x = Bend angle and y = Takeoff velocity.

a. What is the equation of the estimated regression line?
b. What is the estimated average increase in takeoff velocity associated with a 1 degree change in bend angle?
c. What would you predict takeoff velocity to be for a salamander who jumps with a bend angle of 50 degrees?
d. Should you use this regression model to predict takeoff velocity for a salamander who jumps with a bend angle of 90 degrees? Explain.
e. Calculate and interpret the value of r^2.
f. Calculate and interpret the value of s_e.

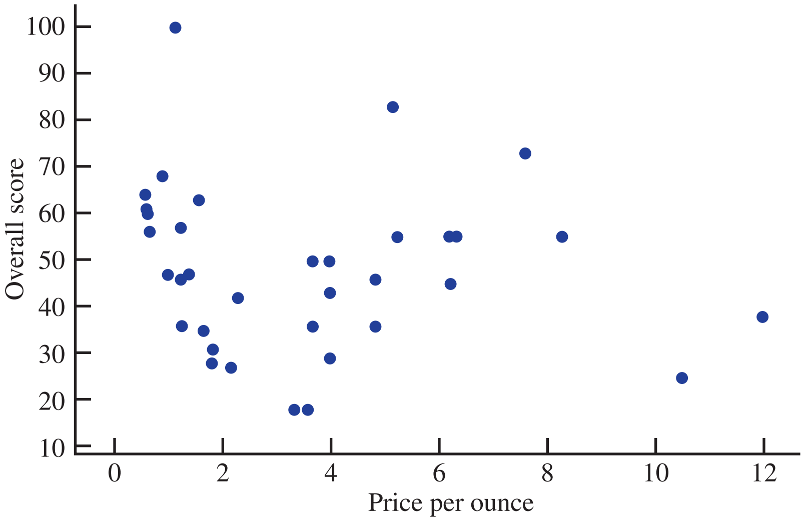
13.11 *Consumer Reports* **(consumerreports.org, July 2023)** studied 35 sunscreen lotions and assigned each one an overall score based on its performance in terms of sun protection, protection against ultraviolet rays, and the difference between the advertised sun protection factor (SPF) and the SPF as measured in laboratory testing. The article also reported the price per ounce (in dollars) for each of the sunscreens. The data in the article were used to produce the accompanying regression output below and scatterplot on the next page.



```
Regression Equation
Overall Score  =  51.11 - 0.88 Price per Ounce

Coefficients
Term             Coef  SE Coef  T-Value  P-Value
Constant        51.11     4.86    10.51    0.000
Price per Ounce -0.88     1.06    -0.83    0.413

Model Summary
      S   R-sq  R-sq(adj)  R-sq(pred)
17.8204  2.04%      0.00%       0.00%
```

Copyright 2025 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s). Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

Output for Exercise 13.11



- **a.** What is the value of r^2? Interpret the value of r^2 in the context of this problem.
- **b.** Explain how the value of r^2 is consistent with what you see in the scatterplot.

13.12 A simple linear regression model was used to describe the relationship between y = Hardness of molded plastic and x = Amount of time elapsed since the end of the molding process. Summary quantities included $n = 15$, SSResid = 1235.470, and SSTo = 25,321.368.

- **a.** Calculate an estimate of σ. What value for degrees of freedom is associated with this estimate?
- **b.** What percentage of observed variation in hardness can be explained by the linear relationship between hardness and elapsed time?

13.13 ● Consider the accompanying data on

x = Advertising share

and

y = Market share

for a particular brand of soft drink during 10 randomly selected years.

x	0.103	0.072	0.071	0.077	0.086	0.047	0.060	0.050	0.070	0.052
y	0.135	0.125	0.120	0.086	0.079	0.076	0.065	0.059	0.051	0.039

- **a.** Construct a scatterplot for these data. Is the simple linear regression model appropriate for describing the relationship between x and y?
- **b.** Calculate the equation of the estimated regression line and use it to obtain the predicted market share when the advertising share is 0.09.
- **c.** Calculate the value of r^2. Interpret this value.
- **d.** Calculate a point estimate of σ. What value for degrees of freedom is associated with this estimate?

Section 13.2 Inferences About the Slope of the Population Regression Line

The slope coefficient β in the simple linear regression model represents the average or expected change in the dependent variable y that is associated with a 1-unit increase in the value of the independent variable x. For example, consider x = the size of a house (in square feet) and y = selling price of the house. If the simple linear regression model is appropriate for the population of houses in a particular city, β would be the average increase in selling price associated with a 1-square foot increase in size. As another example, if x = number of hours per week a computer system is used and y = the annual maintenance expense, then β would be the expected change in expense associated with using the computer system one additional hour per week.

Because the value of β is almost always unknown, it must be estimated from the sample data. The slope b of the least-squares line gives a point estimate. As with any point estimate, it is desirable to have some indication of how accurately b estimates β.

Copyright 2025 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s). Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

In some situations, the value of the statistic b may vary greatly from sample to sample, so b calculated from a single sample may be quite different from the actual population slope β. In other situations, almost all possible samples yield b values that are close to β, so the error of estimation is likely to be small.

In order to make inferences about the slope of the population regression line, you need to know about the sampling distribution of b. In particular, you need to know the shape of the sampling distribution, where the sampling distribution is centered relative to β, and how much it spreads out about its center.

Properties of the Sampling Distribution of *b*

When the four basic assumptions of the simple linear regression model are satisfied, the following statements are true:

1. The mean value of b is β. Because $\mu_b = \beta$, the sampling distribution of b is always centered at the value of β. This means that b is an unbiased statistic for estimating β.
2. The standard deviation of the statistic b is

$$\sigma_b = \frac{\sigma_e}{\sqrt{\sum(x - \bar{x})^2}}$$

3. The statistic b has a normal distribution (this is a consequence of the model assumption that the random deviation e is normally distributed).

The fact that b is unbiased means that the sampling distribution is centered at the right place. The standard deviation of b describes variability in the values of b. If σ_b is large, then the sampling distribution of b will be quite spread out around β and an estimate far from the actual value of β could result. For σ_b to be small, the numerator σ should be small (representing little variability about the population line) and/or the denominator should be large. Because $\Sigma(x - \bar{x})^2$ is a measure of how much the observed x values spread out, β tends to be more precisely estimated when the x values in the sample are spread out rather than when they are close together.

The normality of the sampling distribution of b implies that the standardized variable

$$z = \frac{b - \beta}{\sigma_b}$$

has a standard normal distribution. However, inferential methods cannot be based on this statistic, because the value of σ_b is not known (since the unknown σ_e appears in the numerator of σ_b). One way to proceed is to estimate σ_e with s_e.

The **estimated standard deviation of the statistic *b* is**

$$s_b = \frac{s_e}{\sqrt{\sum(x - \bar{x})^2}}$$

When the four basic assumptions of the simple linear regression model are satisfied, the statistic

$$t = \frac{b - \beta}{s_b}$$

has a t distribution with df $= (n - 2)$.

In the same way that $t = \dfrac{\bar{x} - \mu}{\frac{s}{\sqrt{n}}}$ was used in Chapter 9 to develop a confidence interval for μ, the t variable in the preceding box can be used to obtain a confidence interval for β.

Confidence Interval for β

When the four basic assumptions of the simple linear regression model are satisfied, a **confidence interval for $\boldsymbol{\beta}$**, the slope of the population regression line, has the form

$$b \pm (t \text{ critical value}) \cdot s_b$$

where the t critical value is based on df $= (n - 2)$. Technology or Appendix Table 3 can be used to determine the appropriate t critical value.

The interval estimate of β is centered at b and extends out from the center by an amount that depends on the sampling variability of b. When s_b is small, the interval is narrow, implying relatively precise knowledge of the value of β.

Example 13.4 The Bison of Yellowstone Park

Understand the context ➤

● The dedicated work of conservationists for over 100 years has brought the bison in Yellowstone National Park from near extinction to a herd of over 5900 animals (as of 2022). This recovery is a mixed blessing. Many bison have been exposed to the bacteria that cause brucellosis, a disease that infects domestic cattle, and there are many domestic cattle herds near Yellowstone. Because of concerns that free-ranging bison can infect nearby cattle, it is important to monitor and manage the size of the bison population, and if possible, keep bison from transmitting this bacteria to ranch cattle.

The article **"Reproduction and Survival of Yellowstone Bison"** ***(The Journal of Wildlife Management*** **[2007]: 2365–2372)** describes a large multiyear study of the factors that influence bison movement and herd size. The researchers studied a number of environmental factors to better understand the relationship between bison reproduction and the environment. One factor thought to influence reproduction is stress due to accumulated snow, making foraging more difficult for pregnant bison. Data on

y = Spring calf ratio (SCR)

and

x = Previous fall snow-water equivalent (SWE)

for 17 years are shown in the following table. The spring calf ratio is the ratio of calves to adults, a measure of reproductive success. Snow water equivalent is the depth of water that would result if the snow pack melted. The SWE measurements in the table are in thousands of centimeters (so 1.933 represents 1933 cm). The researchers were interested in estimating the mean change in spring calf ratio associated with each additional 1000 centimeters in snow-water equivalent.

Consider the data ➤

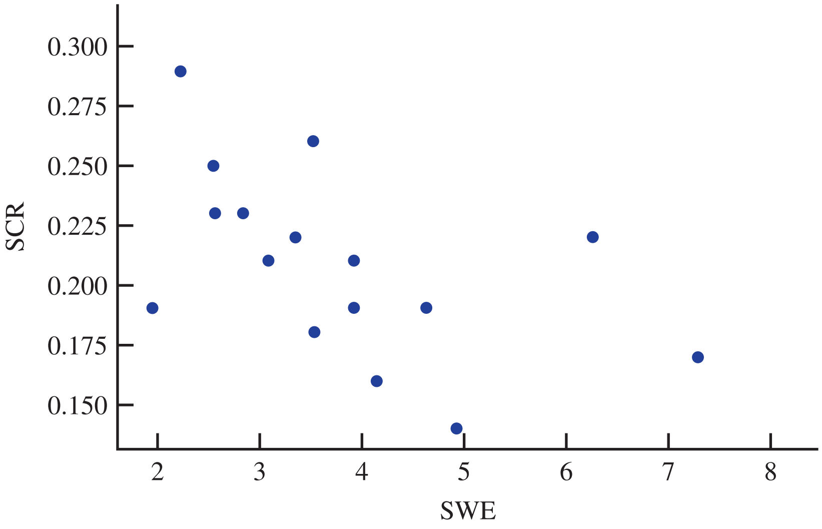
SCR	SWE	SCR	SWE
0.19	1.933	0.22	3.317
0.14	4.906	0.22	3.332
0.21	3.072	0.18	3.511
0.23	2.543	0.21	3.907
0.26	3.509	0.25	2.533
0.19	3.908	0.19	4.611
0.29	2.214	0.22	6.237
0.23	2.816	0.17	7.279
0.16	4.128		

To proceed, you would need to assume that these years are representative of yearly circumstances at Yellowstone, and that each year's reproduction and snowfall is independent of previous years.

● **Data set available online**

Copyright 2025 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s). Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

A scatterplot of the data is shown here.



The plot shows a linear pattern, and the vertical variability of points does not appear to be changing over the range of x values in the sample. If you assume that the distribution of errors at any given x value is approximately normal, then the simple linear regression model would be appropriate. (In Section 13.3, you will see how the assumption that the distribution of errors is normal can be evaluated.).

The slope β in this context is the average change in spring calf ratio associated with each additional 1000 cm in snow water equivalent. The scatterplot shows a negative linear relationship, so the point estimate of β will be negative.

Do the work ➤ Based on the given data

$$n = 17 \qquad \Sigma x = 63.756 \qquad \Sigma y = 3.56$$

$$\Sigma x^2 = 270.425 \quad \Sigma xy = 12.923 \quad \Sigma y^2 = 0.7682$$

from which

$$b = -0.0137 \qquad a = 0.261$$

$$\text{SSResid} = 0.016847 \qquad \text{SSTo} = 0.022694$$

$r^2 = 0.258$ (25.8% of the observed variability in spring calf ratio can be explained by the simple linear regression model)

$$s_e^2 = 0.0011 \qquad s_e = 0.0335$$

$$s_b = \frac{s_e}{\sqrt{\Sigma(x-\bar{x})^2}} = \frac{0.0335}{\sqrt{31.3175}} = 0.006$$

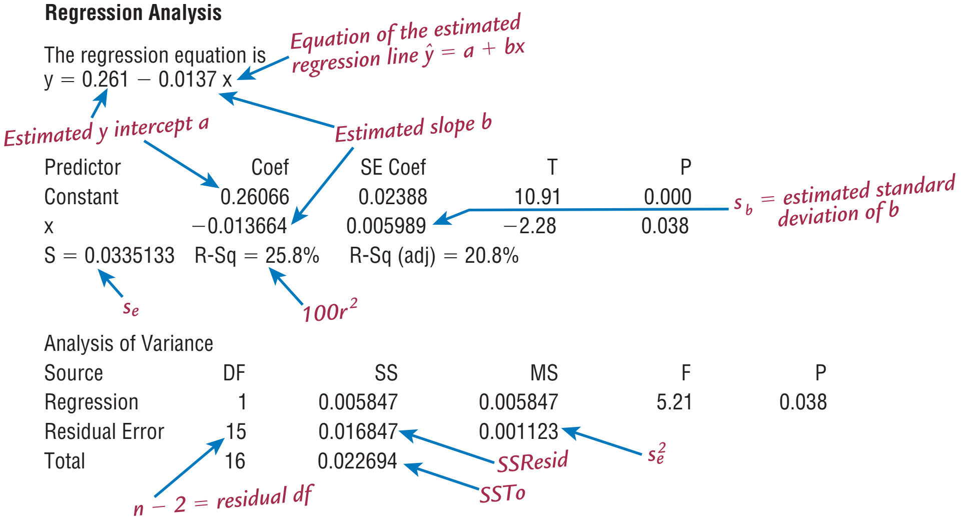
Calculation of the 95% confidence interval for β requires a t critical value based on df $= n - 2 = 17 - 2 = 15$, which (from Appendix Table 3) is 2.13. The resulting interval is then

$$\begin{aligned} b \pm (t \text{ critical value}) \cdot s_b &= -0.0137 \pm (2.13)(0.006) \\ &= -0.0137 \pm 0.0128 \\ &= (-0.027, -0.001) \end{aligned}$$

Interpret the results ➤ You can interpret this interval as follows: Based on the sample data, you are 95% confident that the average change in spring calf ratio associated with a 1000-cm increase in snow-water equivalent is between −0.027 and −0.001. This means that you think that spring calf ratio decreases by somewhere between 0.001 and 0.027 with each additional 1000 cm of snow-water equivalent.

Copyright 2025 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s). Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

Output from any of the standard statistical software packages routinely includes the calculated values of a, b, SSResid, SSTo, and s_b. Figure 13.9 displays partial Minitab output for the data of Example 13.4. The format from other software packages is similar. Rounding occasionally leads to small discrepancies between hand-calculated and computer-calculated values, but there are no such discrepancies in this example.



```
Regression Analysis

The regression equation is
y = 0.261 − 0.0137 x

Predictor        Coef     SE Coef       T       P
Constant      0.26066     0.02388   10.91   0.000
x           −0.013664    0.005989   −2.28   0.038
S = 0.0335133   R-Sq = 25.8%   R-Sq (adj) = 20.8%

Analysis of Variance
Source           DF         SS          MS      F       P
Regression        1   0.005847    0.005847   5.21   0.038
Residual Error   15   0.016847    0.001123
Total            16   0.022694
```

Equation of the estimated regression line $\hat{y} = a + bx$
Estimated y intercept a
Estimated slope b
s_b = estimated standard deviation of b
s_e
$100r^2$
SSResid
SSTo
s_e^2
$n - 2$ = residual df

Figure 13.9 Partial Minitab output for the data of Example 13.4.

Hypothesis Tests Concerning β

Hypotheses about β can be tested using a t test similar to the t tests introduced in Chapters 10 and 11. The null hypothesis states that β has a specified hypothesized value. The t statistic results from standardizing b, the point estimate of β, under the assumption that H_0 is true. When H_0 is true, the sampling distribution of this statistic is the t distribution with df $= (n - 2)$.

Frequently, the null hypothesis of interest is $\beta = 0$. If this is the case, the population regression line is a horizontal line, and the value of y in the simple linear regression model does not depend on x. That is,

$$y = \alpha + 0 \cdot x + e = \alpha + e$$

Hypothesis Tests Concerning β

Null Hypothesis: H_0: β = hypothesized value

Test Statistic: $t = \dfrac{b - \text{hypothesized value}}{s_b}$

The test is based on df $= (n - 2)$.

Alternative Hypothesis:	***P*-value:**
H_a: $\beta >$ hypothesized value	Area to the right of the calculated t under the appropriate t curve
H_a: $\beta <$ hypothesized value	Area to the left of the calculated t under the appropriate t curve
H_a: $\beta \neq$ hypothesized value	(1) 2(area to the right of t) if t is positive or (2) 2(area to the left of the t) if t is negative

(continued)

Copyright 2025 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s). Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

Assumptions: For this test to be appropriate, the four basic assumptions of the simple linear regression model must be satisfied:

1. The distribution of e at any particular x value has mean value 0 (that is $\mu_e = 0$).
2. The standard deviation of e is σ, which does not depend on x.
3. The distribution of e at any particular x value is normal.
4. The random deviations $e_1, e_2, \ldots, e_n$ associated with different observations are independent of one another.

What Do You Think?

Why do you think the test of

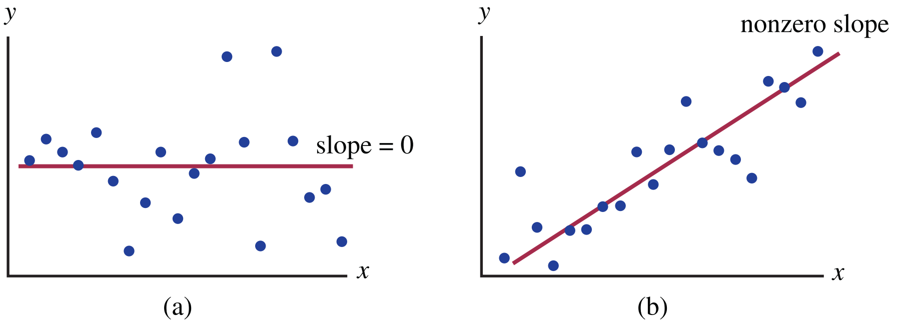
$$H_0 : \beta = 0 \text{ versus } H_a : \beta \neq 0$$

is called the "model utility" test?

When $\beta = 0$, knowledge of x is not useful for predicting y. On the other hand, if $\beta \neq 0$, there is a linear relationship between x and y, and knowledge of x is useful for predicting y. This is illustrated by the scatterplots in Figure 13.10.

The test of H_0: $\beta = 0$ versus H_a: $\beta \neq 0$ is called the **model utility test for simple linear regression**.

Figure 13.10
(a) $\beta = 0$; (b) $\beta \neq 0$.



Model Utility Test for Simple Linear Regression

The **model utility test for simple linear regression** is the test of

$$H_0 : \beta = 0$$

versus

$$H_a : \beta \neq 0$$

The null hypothesis specifies that there is *no* useful linear relationship between x and y, whereas the alternative hypothesis specifies that there *is* a useful linear relationship between x and y. If H_0 is rejected, you would conclude that the simple linear regression model is useful for predicting y.

The test statistic is

$$t = \frac{b - 0}{s_b} = \frac{b}{s_b}.$$

Copyright 2025 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s). Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

If a scatterplot and the r^2 value do not provide clear evidence for a useful linear relationship, it is recommended that the model utility test be carried out before using the regression line to make predictions.

Example 13.5 University Graduation Rates

● The accompanying data on 6-year graduation rate (%), student-related expenditure per full-time student, and median SAT score for a random sample of the primarily undergraduate public universities and colleges in the United States with enrollments between 10,000 and 20,000 were taken from **College Results Online, The Education Trust**.

Median SAT	Expenditure	Graduation Rate
1065	7970	49
950	6401	33
1045	6285	37
990	6792	49
950	4541	22
970	7186	38
980	7736	39
1080	6382	52
1035	7323	53
1010	6531	41
1010	6216	38
930	7375	37
1005	7874	45
1090	6355	57
1085	6261	48

Start by investigating the relationship between graduation rate and median SAT score. With y = Graduation rate and x = Median SAT score, summary statistics are as follows:

$$n = 15 \qquad \Sigma x = 15{,}195 \qquad \Sigma y = 638$$
$$\Sigma x^2 = 15{,}430{,}725 \qquad \Sigma xy = 651{,}340 \qquad \Sigma y^2 = 28{,}294$$

from which

$$b = 0.132 \qquad a = -91.31 \qquad \text{SSResid} = 491.01$$
$$s_e = 6.146 \qquad r^2 = 0.576 \qquad \Sigma(x - \bar{x})^2 = 38{,}190$$

Because $r^2 = 0.576$, about 57.6% of observed variation in graduation rates can be explained by the simple linear regression model. This suggests that there is a useful linear relation between the two variables, and this can be confirmed by the model utility test. A significance level of 0.05 will be used to carry out this test.

Understand the context ➤

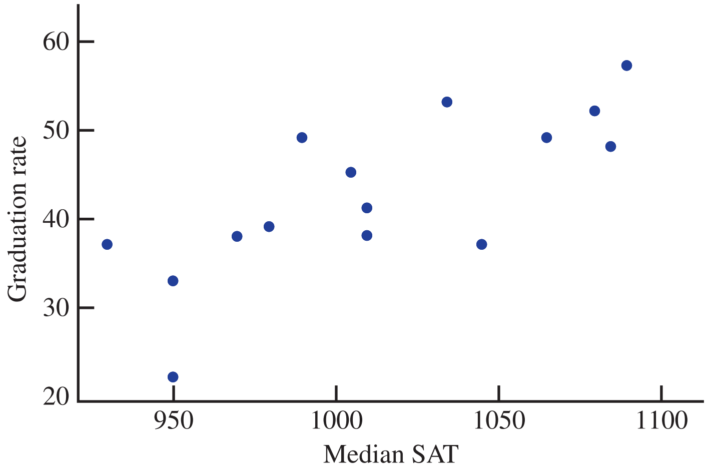
1. β = the average change in graduation rate associated with an increase of 1 unit in median SAT score
2. H_0: $\beta = 0$
3. H_a: $\beta \neq 0$
4. $\alpha = 0.05$

Formulate a plan ➤

5. Test statistic: $t = \dfrac{b - \text{hypothesized value}}{s_b} = \dfrac{b - 0}{s_b} = \dfrac{b}{s_b}$
6. Assumptions: The data are from a random sample, so the observations are independent. The accompanying scatterplot of the data shows a linear pattern and the vertical variability of points does not appear to be changing with x:

● **Data set available online**

Copyright 2025 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s). Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.



Assuming that the distribution of errors at any given x value is approximately normal, the assumptions of the simple linear regression model are satisfied.

Do the work ➤

7. Calculation: The calculation of t requires

$$s_b = \frac{s_e}{\sqrt{\sum(x - \bar{x})^2}} = \frac{6.146}{195.423} = 0.031$$

yielding

$$t = \frac{0.132 - 0}{0.031} = 4.26$$

8. P-value: Appendix Table 4 shows that for a t test with 13 df, $P(t > 4.26) < 0.001$. The inequality in H_a specifies a two-tailed test, so

$$P\text{-value} < 2(0.001) = 0.002.$$

Interpret the results ➤

9. Conclusion: Since P-value < 0.002 is smaller than the significance level 0.05, H_0 is rejected. You can conclude that there is a useful linear relationship between graduation rate and median SAT score.

Figure 13.11 shows partial Minitab output from a simple linear regression analysis. The Coef column gives $b = 0.13213$; $s_b = 0.03145$ is in the SE Coef column; the T column (for t ratio) contains the value of the test statistic for testing H_0: $\beta = 0$; and the P-value for the model utility test is given in the last column as 0.001 (slightly different from the values given in Step 8 because of rounding and because the use of the table produces only approximate P-values). Other commonly used statistical software packages also include this information in their output.

Figure 13.11 Minitab output for the data of Example 13.5.

```
Regression Analysis: Graduation Rate versus Median SAT

The regression equation is
Graduation Rate = –91.3 + 0.132 Median SAT


Predictor          Coef     SE Coef       T        P
Constant         –91.31       31.90   –2.86    0.013
Median SAT      0.13213     0.03145    4.20    0.001


S = 6.14574    R-Sq = 57.6%    R-Sq(adj) = 54.3%


Analysis of Variance

Source            DF        SS        MS        F        P
Regression         1    666.72    666.72    17.65    0.001
Residual Error    13    491.01     37.77
Total             14   1157.73
```

Next, consider the relationship between graduation rate and expenditure per full-time student. Figure 13.12 shows partial Minitab output from a simple linear regression with expenditure as the predictor variable.

Figure 13.12
Minitab output using expenditure as the predictor.

```
The regression equation is
Graduation Rate = 10.9 + 0.00468 Expenditure


Predictor                Coef       SE Coef       T        P
Constant                10.95         17.51    0.63    0.543
Expenditure          0.004680      0.002574    1.82    0.092


S = 8.42608     R-Sq = 20.3%     R-Sq(adj) = 14.1%
```

The value of the test statistic for the model utility test in this case is $t = 1.82$ and the associated P-value is 0.092. For a 0.05 level of significance, you would not reject the null hypothesis of H_0: $\beta = 0$. There is not convincing evidence of a linear relationship between graduation rate and expenditure per full-time student.

When H_0: $\beta = 0$ cannot be rejected by the model utility test at a reasonably small significance level, other models might be considered. One possibility is to use a nonlinear model—an appropriate strategy if the scatterplot shows curvature. Alternatively, a multiple regression model using more than one predictor variable could be considered. Multiple regression models are introduced in Chapter 14.

Exercises 13.14–13.30

● **Data set available online**

13.14 What is the difference between σ and σ_b? What is the difference between σ_b and s_b?

13.15 The largest commercial fishing enterprise in the southeastern United States is the harvest of shrimp. In a study described in the paper **“Long-term Trawl Monitoring of White Shrimp, *Litopenaeus setiferus* (Linnaeus), Stocks within the ACE Basin National Estuariene Research Reserve, South Carolina” (*Journal of Coastal Research* [2008]: 193–199)**, researchers monitored variables thought to be related to the abundance of white shrimp.

One variable the researchers thought might be related to abundance is the amount of oxygen in the water. The relationship between the mean number of white shrimp caught in a single outing and oxygen saturation was described by fitting a regression line using data from 10 randomly selected offshore sites. Computer output is shown below.

```
The regression equation is
Mean number caught = - 5859 + 97.2 Saturation
Predictor          Coef   SE Coef       T      P
Constant          -5859      2394   -2.45  0.040
Saturation        97.22     34.63    2.81  0.023
S = 481.632   R-Sq = 49.6%   R-Sq(adj) = 43.3%
```

a. Is there convincing evidence of a useful linear relationship between the mean number of shrimp caught and oxygen saturation? Explain. (Hint: See Example 13.5.)

b. Is the relationship strong? Explain why or why not.

c. Construct a 95% confidence interval for β and interpret it in context. (Hint: See Example 13.4.)

13.16 Refer back to Example 13.3 in which the simple linear regression model was fit to data on x = Proportion who judged candidate A as more competent and y = Vote difference proportion. For the purpose of estimating β as accurately as possible, would it have been preferable to have observations with x values 0.05, 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9, 0.95, and 0.98? Explain.

13.17 Exercise 13.12 summarized data on y = Hardness of molded plastic and x = Time elapsed since the molding was completed.

Summary quantities included

$$n = 15 \qquad b = 2.50 \qquad \text{SSResid} = 1235.470$$

$$\Sigma(x - \bar{x})^2 = 4024.20$$

a. Calculate the estimated standard deviation of the statistic b.

b. Calculate a 95% confidence interval for β, the slope of the population regression line. (Hint: See Example 13.4.)

c. Does the interval in Part (b) suggest that β has been precisely estimated? Explain.

13.18 A simple linear regression model was used to describe the relationship between sales revenue y (in thousands of dollars) and advertising expenditure x (also in thousands of dollars) for fast-food outlets during a 3-month period. A random sample of 15 outlets resulted in the accompanying summary quantities.

$$\Sigma x = 14.10 \quad \Sigma y = 1438.50 \quad \Sigma x^2 = 13.92$$
$$\Sigma y^2 = 140{,}354 \quad \Sigma xy = 1387.20$$
$$\Sigma(y - \bar{y})^2 = 2401.85 \quad \Sigma(y - \hat{y})^2 = 561.46$$

a. What proportion of observed variation in sales revenue can be attributed to the linear relationship between revenue and advertising expenditure?
b. Calculate s_e and s_b.
c. Calculate a 90% confidence interval for β, the average change in revenue associated with a \$1000 (that is, 1 unit) increase in advertising expenditure.

13.19 An experiment to study the relationship between x = Time spent exercising (minutes) and y = Amount of oxygen consumed during the exercise period resulted in the following summary statistics.

$$n = 20 \quad \Sigma x = 50 \quad \Sigma y = 16{,}705 \quad \Sigma x^2 = 150$$
$$\Sigma y^2 = 14{,}194{,}231 \quad \Sigma xy = 44{,}194$$

a. Estimate the slope and y intercept of the population regression line.
b. One sample observation on oxygen usage was 757 for a 2-minute exercise period. What amount of oxygen consumption is predicted for this exercise period, and what is the corresponding residual?
c. Calculate a 99% confidence interval for the average change in oxygen consumption associated with a 1-minute increase in exercise time.

13.20 The paper **"The Effects of Split Keyboard Geometry on Upper Body Postures" (*Ergonomics* [2009]: 104–111)** describes a study to determine the effects of several keyboard characteristics on typing speed. One of the variables considered was the front-to-back surface angle of the keyboard. Minitab output resulting from fitting the simple linear regression model with x = Surface angle (degrees) and y = Typing speed (words per minute) is given below.

```
Regression Analysis: Typing Speed versus
Surface Angle
The regression equation is
Typing Speed = 60.0 + 0.0036 Surface Angle

Predictor         Coef    SE Coef       T      P
Constant       60.0286     0.2466  243.45  0.000
Surface        0.00357    0.03823    0.09  0.931
Angle

S = 0.511766   R-Sq = 0.3%   R-Sq(adj) = 0.0%

Analysis of Variance
Source           DF      SS      MS     F      P
Regression        1  0.0023  0.0023  0.01  0.931
Residual Error    3  0.7857  0.2619
Total             4  0.7880
```

a. Assuming that the basic assumptions of the simple linear regression model are reasonably met, carry out a hypothesis test to decide if there is a useful linear relationship between x and y. (Hint: See Example 13.5.)
b. Are the values of s_e and r^2 consistent with the conclusion from Part (a)? Explain.

13.21 The authors of the paper **"Decreased Brain Volume in Adults with Childhood Lead Exposure" (*Public Library of Science Medicine* [May 27, 2008]: e112)** studied the relationship between childhood environmental lead exposure and a measure of brain volume change in a particular region of the brain. Data were given for x = Mean childhood blood lead level (μg/dL) and y = Brain volume change (BVC, in percent). A subset of data read from a graph that appeared in the paper was used to produce the accompanying Minitab output.

```
Regression Analysis: BVC versus Mean Blood
Lead Level
The regression equation is
BVC = -0.00179 - 0.00210 Mean Blood Lead Level

Predictor              Coef   SE Coef      T     P
Constant          -0.001790  0.008303  -0.22 0.830
Mean Blood Lead -0.0021007 0.0005743  -3.66 0.000
Level
```

Carry out a hypothesis test to decide if there is convincing evidence of a useful linear relationship between x and y. Assume that the basic assumptions of the simple linear regression model are reasonably met.

13.22 Forensic scientists sometimes use the length of various bones to estimate the height of the individual from which the bone was recovered. The online article **"What Relationship Exists Between Height and Forearm Length?" (heightcomparison.com, March 31, 2021)** summarizes several studies that investigated if forearm length is a good predictor of height. One of these studies used a linear regression model to describe the relationship between x = Forearm length (in cm) and y = Height (in cm). The estimated regression line was $\hat{y} = 73.69 + 3.54x$. Suppose that the estimated regression equation was calculated using data from $n = 100$ people and that $s_b = 0.6$. Carry out a hypothesis test to decide if there is convincing evidence of a useful linear relationship between forearm length and height. For purposes of this exercise, you can assume that the basic assumptions of the simple linear regression model are met.

13.23 ● Researchers studying pleasant touch sensations measured the firing frequency (impulses per second) of nerves that were stimulated by a light brushing stroke on the forearm and also recorded the subject's numerical rating of how pleasant the sensation was. The accompanying data was read from a graph in the paper **"Coding of Pleasant Touch by Unmyelinated Afferents in Humans" (*Nature Neuroscience,* April 12, 2009)**.

Firing Frequency	**Pleasantness Rating**
23	0.2
24	1.0
22	1.2
25	1.2
27	1.0
28	2.0
34	2.3
33	2.2
36	2.4
34	2.8

a. Estimate the mean change in pleasantness rating associated with an increase of 1 impulse per second in firing frequency using a 95% confidence interval. Interpret the resulting interval.

b. Carry out a hypothesis test to decide if there is convincing evidence of a useful linear relationship between firing frequency and pleasantness rating.

13.24 ● The data in the accompanying table are a subset of data from the report **"Great Jobs, Great Lives" (Gallup-Purdue Index 2015 Report, gallup.com)**. The data are approximate values read from a scatterplot. University students were asked if they agreed that their education was worth the cost. One variable in the table is the *U.S. News and World Report* ranking of the university in 2015. The other variable in the table is the percentage of students at the university who responded "strongly agree."

University Ranking	**Percentage of Alumni Who Strongly Agree**
28	53
29	58
30	62
37	55
45	54
47	62
52	55
54	62
57	70
60	58
65	66
66	55
72	65
75	57
82	67
88	59
98	75

(continued)

a. Fit a linear regression model to predict the percentage of alumni who strongly agree that their education was worth the cost, using 2015 university ranking as a predictor.

b. Do the sample data support the hypothesis that there is a useful linear relationship between the percentage of alumni who strongly agree that their education was worth the cost and 2015 university ranking? Test the appropriate hypotheses using $\alpha = 0.05$.

13.25 ● Acrylamide is a chemical that is sometimes found in cooked starchy foods and which is thought to increase the risk of certain kinds of cancer. The paper **"A Statistical Regression Model for the Estimation of Acrylamide Concentrations in French Fries for Excess Lifetime Cancer Risk Assessment" (*Food and Chemical Toxicology* [2012]: 3867–3876)** describes a study to investigate the effect of frying time (in seconds) and acrylamide concentration (in micrograms per kilogram) in French fries. The data in the accompanying table are approximate values read from a graph that appeared in the paper.

Frying Time	**Acrylamide Concentration**
150	155
240	120
240	190
270	185
300	140
300	270

a. For these data, the estimated regression line for predicting $y =$ Acrylamide concentration based on $x =$ Frying time is $\hat{y} = 87 + 0.359x$. What is an estimate of the average change in acrylamide concentration associated with a 1-second increase in frying time?

b. What would you predict for acrylamide concentration for a frying time of 250 seconds?

c. Use the given Minitab on the next page to decide if there is convincing evidence of a useful linear relationship between acrylamide concentration and frying temperature.

```
Regression Equation
y = 87 + 0.359x
```

```
Coefficients
Term          Coef  SE Coef  T-Value  P-Value    VIF
Constant        87      112     0.78    0.480
x            0.359    0.438     0.82    0.459   1.00

      S      R-sq  R-sq(adj)  R-sq(pred)
54.7108    14.38%      0.00%       0.00%
```

13.26 The paper referenced in Exercise 13.9 (**"Depression, Body Mass Index, and Chronic Obstructive Pulmonary Disease—A Holistic Approach,"** ***International Journal of COPD*** **[2016]: 239–249)** gave data on

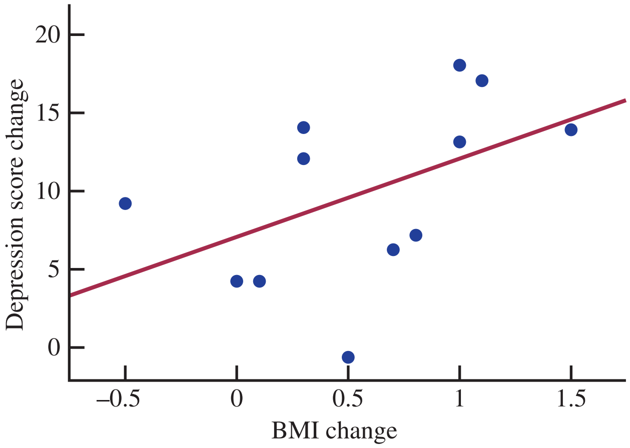
x = Change in body mass index (BMI in kg/m^2)

and

y = Change in a measure of depression

for patients suffering from depression who participated in a pulmonary rehabilitation program. JMP output for these data is shown below.

Bivariate Fit of Depression score change By BMI Change



```
Linear Fit
Depression score change = 6.8725681 +
5.077821*BMI Change
Summary of Fit
RSquare                                0.234828
RSquare Adj                            0.158311
Root Mean Square Error                 5.365593
Mean of Response                           9.75
Observations (or Sum Wgts)                   12

Analysis of Variance
Source     DF Sum of Squares Mean Square  F Ratio
Model       1       88.35409     88.3541   3.0690
Error      10      287.89591     28.7896  Prob > F
C. Total   11      376.25000               0.1104
```

```
Parameter Estimates
Term            Estimate Std Error t Ratio Prob>|t|
Intercept      6.8725681  2.257651    3.04  0.0124*
BMI Change      5.077821  2.898557    1.75  0.1104
```

a. What does the scatterplot suggest about the relationship between depression score change and BMI change?

b. What is the equation of the estimated regression line?

c. Is there a useful linear relationship between the two variables? Carry out an appropriate test using a significance level of $\alpha = 0.05$.

13.27 Exercise 13.18 described a regression analysis with y = Sales revenue and x = Advertising expenditure. Summary quantities given there result in

$n = 15 \qquad b = 52.27 \qquad s_b = 8.05$

a. Test the hypothesis H_0: $\beta = 0$ versus H_a: $\beta \neq 0$ using a significance level of 0.05. What does the conclusion say about the nature of the relationship between x and y?

b. Consider the hypothesis H_0: $\beta = 40$ versus H_a: $\beta > 40$. The null hypothesis states that the average change in sales revenue associated with a 1-unit increase in advertising expenditure is (at most) \$40,000. Carry out a test using significance level 0.01.

13.28 ● Consider the accompanying data on x = Research and development expenditure (thousands of dollars) and y = Growth rate (% per year) for eight different industries.

x	2,024	5,038	905	3,572	1,157	327	378	191
y	1.90	3.96	2.44	0.88	0.37	−0.90	0.49	1.01

a. Would a simple linear regression model provide useful information for predicting growth rate from research and development expenditure? Test the appropriate hypotheses using a 0.05 significance level.

b. Use a 90% confidence interval to estimate the average change in growth rate associated with a \$1000 increase in expenditure. Interpret the resulting interval.

13.29 Suppose that a single y observation is made at each of the x values 5, 10, 15, 20, and 25.

a. If $\sigma = 4$, what is the standard deviation of the statistic b?

b. Now suppose that a second observation is made at every x value listed in Part (a) (for a total of 10 observations). Is the resulting value of σ_b half of what it was in Part (a)?

c. How many observations at each x value in Part (a) are required to yield a σ_b value that is half the value calculated in Part (a)?

Copyright 2025 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s). Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

13.30 ● In anthropological studies, an important characteristic of fossils is cranial capacity. Frequently skulls are at least partially decomposed, so it is necessary to use other characteristics to obtain information about cranial capacity. One measure that has been used is the length of the lambda-opisthion chord. The article **"Vertesszollos and the Presapiens Theory" (*American Journal of Physical Anthropology* [1971])** reported the accompanying data for $n = 7$ *Homo erectus* fossils.

x (Chord length in mm)	78	75	78	81	84	86	87
y (Capacity in cm^3)	850	775	750	975	915	1,015	1,030

Suppose that from previous evidence, anthropologists had believed that for each 1-mm increase in chord length, cranial capacity would be expected to increase by 20 cm^3. Do the data given here strongly contradict prior belief?

Section 13.3 Checking Model Adequacy

The simple linear regression model equation is

$$y = \alpha + \beta x + e$$

where e represents the random deviation of an observed y value from the population regression line $\alpha + \beta x$. The inferential methods presented in Section 13.2 required some assumptions about e, including these:

1. At any particular x value, the distribution of e is a normal distribution.
2. At any particular x value, the standard deviation of e is σ, which is the same for all values of x (meaning that σ does not depend on x).

Inferences based on the simple linear regression model continue to be reliable when model assumptions are slightly violated (for example, mild nonnormality of the random deviation distribution). However, using an estimated model in the face of substantially violated assumptions can result in misleading conclusions. In this section, methods for identifying such serious violations of model assumptions are considered.

Residual Analysis

If the deviations $e_1, e_2, \ldots, e_n$ from the population line were available, they could be examined for any inconsistencies with model assumptions. For example, a normal probability plot would suggest whether or not normality was reasonable. But, because

$$\begin{aligned} e_1 &= y_1 - (\alpha + \beta x_1) \\ &\vdots \\ e_n &= y_n - (\alpha + \beta x_n) \end{aligned}$$

these deviations can be calculated only if the equation of the population line is known. In practice, this is not the case. Instead, diagnostic checks are based on the residuals

$$\begin{aligned} y_1 - \hat{y}_1 &= y_1 - (a + bx_1) \\ &\vdots \\ y_n - \hat{y}_n &= y_n - (a + bx_n) \end{aligned}$$

which are the deviations from the *estimated* line.

When all model assumptions are met, the mean value of the residuals at any particular x value is 0. Any observation that gives a large positive or negative residual should be examined carefully for any unusual circumstances, such as a recording error or exceptional experimental conditions. Identifying residuals with unusually large magnitudes is made easier by considering the **standardized residuals**.

Recall that a quantity is standardized by subtracting its mean value (0 in this case) and dividing by its estimated standard deviation. So, to obtain standardized residuals, you would calculate

$$\text{standardized residual} = \frac{\text{residual}}{\text{estimated standard deviation of residual}}$$

What Do You Think?

How would you interpret a standardized residual of 0.5? How would you interpret a standardized residual of 3.1?

The value of a standardized residual tells you how many standard deviations the corresponding residual is from its expected value, 0.

Because residuals at different x values have different standard deviations* (depending on the value of x for that observation), calculating the standardized residuals can be tedious. Fortunately, many statistical software packages provide standardized residuals as part of the output.

In Chapter 7, the normal probability plot was introduced as a technique for deciding whether the n observations in a random sample could plausibly have come from a normal population distribution. To assess whether the assumption that $e_1, e_2, \ldots e_n$ all come from the same normal distribution is reasonable, you can construct a normal probability plot of the standardized residuals. This is illustrated in the following example.

Example 13.6 Political Faces

Understand the context ➤

● Example 13.3 introduced data on

x = Proportion who judged candidate A as the more competent of two candidates based on facial appearance

and

y = Vote difference (candidate A − candidate B) expressed as a proportion of the total number of votes cast

for a sample of 12 congressional elections. (See Example 13.3 for a more detailed description of the study.)

The scatterplot in Figure 13.13 is consistent with the assumptions of the simple linear regression model.

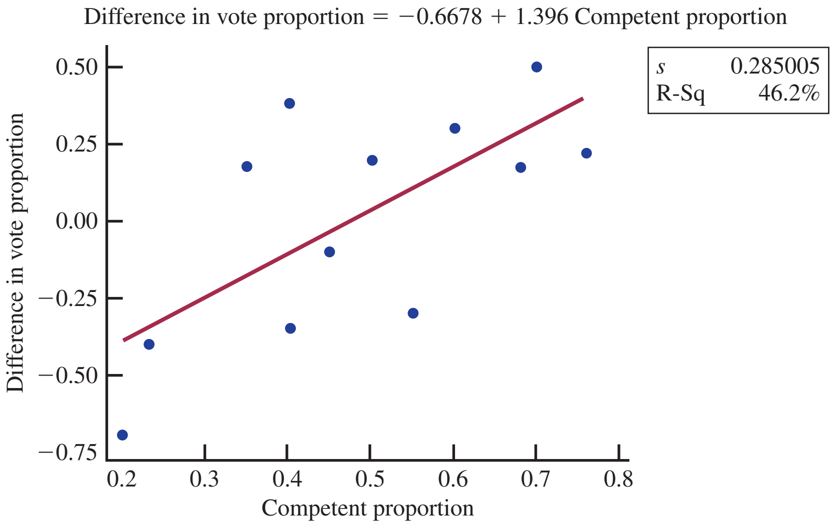


Figure 13.13 Minitab output for the data of Example 13.6.

● **Data set available online**

---

*The estimated standard deviation of the ith residual, $(y_i - \hat{y}_i)$ is $s_e\sqrt{1 - \frac{1}{n} - \frac{(x_i - \bar{x})^2}{S_{xx}}}$.

Copyright 2025 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s). Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

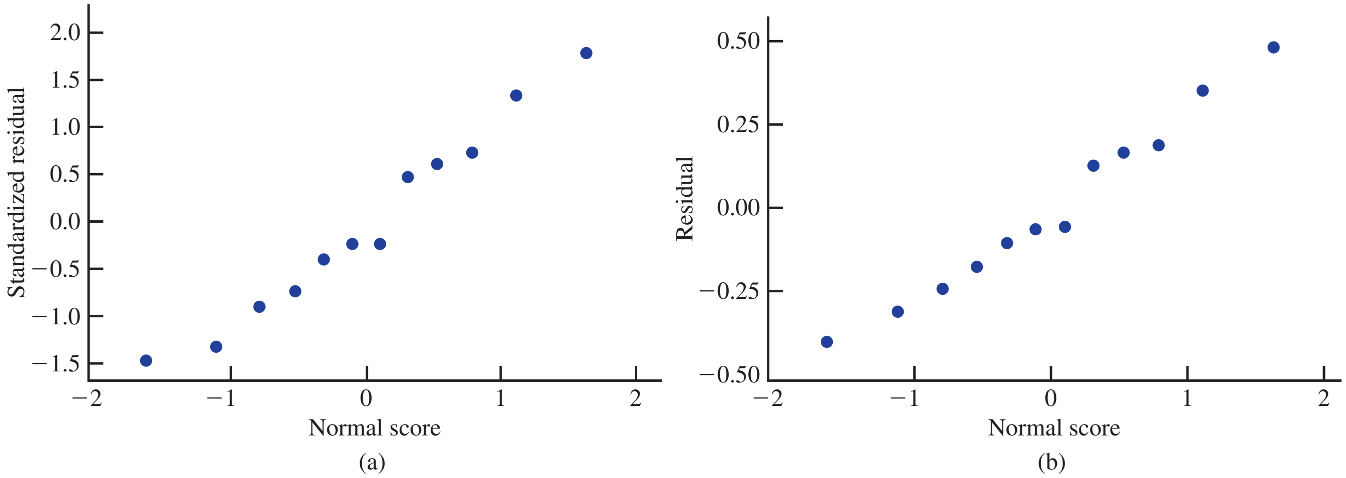
The residuals, their standard deviations, and the standardized residuals (computed using Minitab) are given in Table 13.1. For the residual with the largest magnitude, 0.49, the standardized residual is 1.81. This means that this residual is approximately 1.8 standard deviations above its expected value of 0, which is not particularly unusual in a sample of this size. On the standardized scale, no residual here is surprisingly large.

Table 13.1 **Data, Residuals, and Standardized Residuals for Example 13.6**

Observation	Competent Proportion x	Difference in Vote Proportion y	$\hat{y}$	Residual	Estimated Standard Deviation of Residual	Standardized Residual
1	0.20	−0.70	−0.39	−0.31	0.24	−1.32
2	0.23	−0.40	−0.35	−0.05	0.24	−0.22
3	0.40	−0.35	−0.11	−0.24	0.27	−0.89
4	0.35	0.18	−0.18	0.36	0.27	1.35
5	0.40	0.38	−0.11	0.49	0.27	1.81
6	0.45	−0.10	−0.04	−0.06	0.27	−0.22
7	0.50	0.20	0.03	0.17	0.27	0.62
8	0.55	−0.30	0.10	−0.40	0.27	−1.48
9	0.60	0.30	0.17	0.13	0.27	0.49
10	0.68	0.18	0.28	−0.10	0.26	−0.39
11	0.70	0.50	0.31	0.19	0.25	0.75
12	0.76	0.22	0.39	−0.17	0.24	−0.72

Figure 13.14 displays a normal probability plot of the standardized residuals and also a separate normal probability plot of the residuals. Notice that in this case the plots are nearly identical. It is usually the case that the two plots are similar. Although it is preferable to work with the standardized residuals, without access to a statistical software package or calculator that will produce standardized residuals, a plot of the unstandardized residuals can be used. Both of the normal probability plots in Figure 13.14 are approximately linear. The plots would not cause you to question the assumption of normality.

Figure 13.14
Normal probability plots for Example 13.6 (from Minitab): (a) standardized residuals; (b) residuals.



(a)

(b)

In addition to a normal probability plot, a boxplot of the residuals might also be used to assess whether the assumption of normality is reasonable. A boxplot that is reasonably symmetric and that does not show any outliers is consistent with the assumption of normality.

Plotting the Residuals

A plot of the (x, residual) pairs is called a **residual plot**, and a plot of the (x, standardized residual) pairs is a **standardized residual plot**. Residual and standardized residual plots typically exhibit the same general shapes. With access to a statistical software package or graphing calculator that calculates standardized residuals, using the standardized residual plot is recommended. If this is not possible, it is acceptable to use the residual plot instead.

A standardized residual plot or a residual plot is often helpful in identifying unusual or highly influential observations and in checking for violations of model assumptions. A desirable plot is one that exhibits no particular pattern (such as curvature or a much greater vertical variability in one part of the plot than in another) and that has no point that is far removed from all the others.

A point falling far above or far below the horizontal line *standardized residual* $= 0$ corresponds to a large standardized residual, which can indicate some kind of unusual behavior, such as a recording error, a nonstandard experimental condition, or an unusual experimental subject. A point that has an x value that differs greatly from others in the data set could have exerted excessive influence in determining the fitted line.

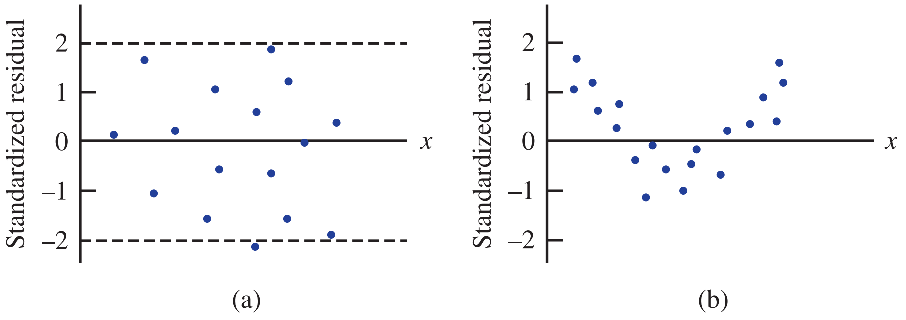
A standardized residual plot like the one pictured in Figure 13.15(a) is desirable, because no point lies much outside the horizontal band between -2 and 2 (so there is no unusually large standardized residual corresponding to an outlying observation). There is also no point far to the left or right of the others (which could indicate an observation that might greatly influence the fit). Finally there is no pattern to indicate that the model should somehow be modified.

When the plot has the appearance of Figure 13.15(b), the fitted model should be changed to incorporate curvature (a nonlinear model). The increasing variability from left to right in Figure 13.15(c) suggests that the variance of y is not the same at each x value but rather increases with x. A linear model may still be appropriate, but the best-fit line should be obtained by using *weighted least* squares rather than ordinary least squares. This involves giving more weight to observations in the region exhibiting low variability and less weight to observations in the region exhibiting high variability. A specialized regression analysis textbook or a statistician should be consulted for more information on using weighted least squares.

The standardized residual plots of Figures 13.15(d) and 13.15(e) show an **outlier** (a point with a large standardized residual) and a **potentially influential observation**, respectively. Consider deleting the potentially influential observation from the data set and refitting the same model. Substantial changes in estimates and other summary quantities are a signal that a more careful analysis should be carried out before proceeding.

Figure 13.15

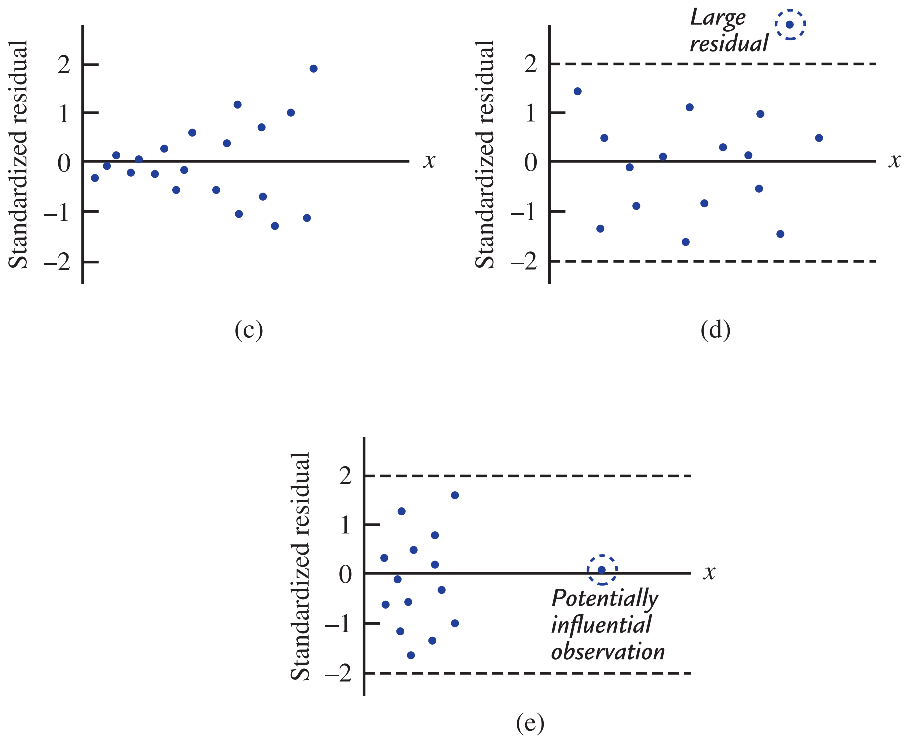
Examples of residual plots: (a) satisfactory plot; (b) plot suggesting that a curvilinear regression model is needed



(a) (b)

(continued)

Figure 13.15 continued
Examples of residual plots: (c) plot indicating nonconstant variance; (d) plot showing a large residual; (e) plot showing a potentially influential observation.



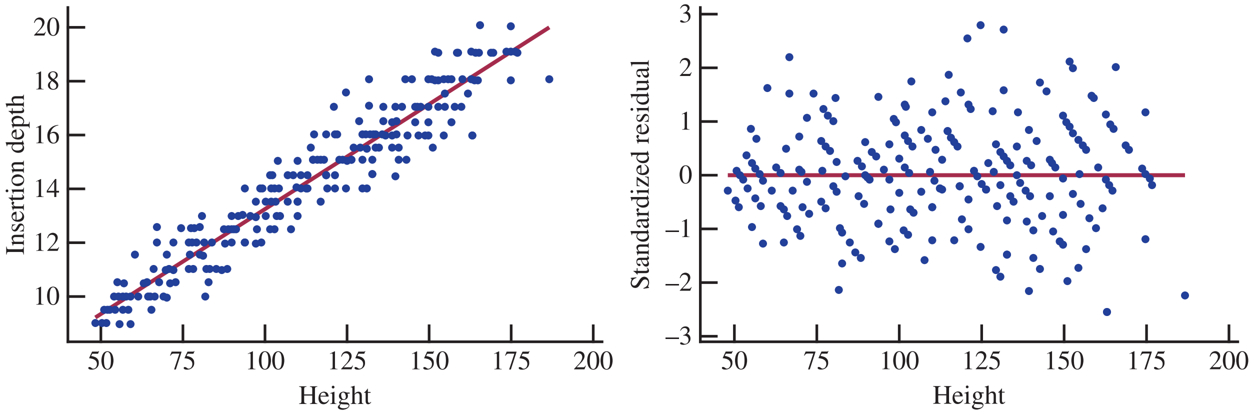
(c) (d) (e)

Example 13.7 A New Pediatric Tracheal Tube

Understand the context ➤

The article **“Appropriate Placement of Intubation Depth Marks in a New Cuffed, Paediatric Tracheal Tube” (*British Journal of Anaesthesia* [2004]: 80–87)** describes a study of the use of tracheal tubes in newborns and infants. Newborns and infants have small trachea, and there is little margin for error when inserting tracheal tubes. Using X-rays of a large number of children age 2 months to 14 years, the researchers examined the relationships between appropriate trachea tube insertion depth and other variables such as height, weight, and age. A scatterplot and a standardized residual plot constructed using data on the insertion depth and height of the children (both measured in cm) are shown in Figure 13.16.

Figure 13.16
(a) Scatterplot for insertion depth vs. height data of Example 13.7; (b) standardized residual plot.



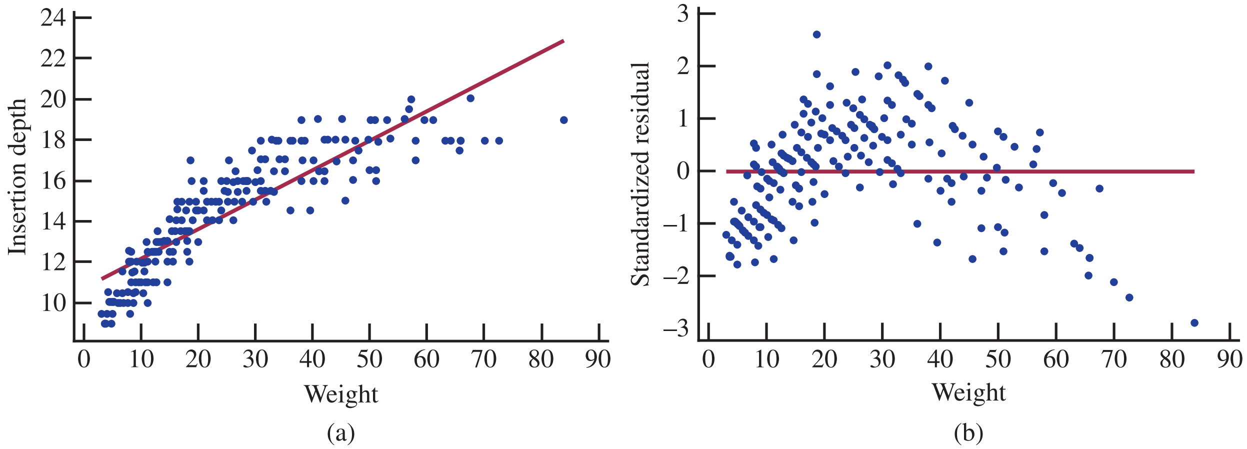
(a) (b)

Copyright 2025 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s). Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

Residual plots like the ones pictured in Figure 13.16(b) are desirable. No point lies outside the horizontal band between −3 and 3, and most are between −2 and 2 (so there are no unusually large residuals corresponding to outliers). There is no point far to the left or right of the others (that is, there are no observations that might be influential), and there is no pattern of curvature or differences in the variability of the residuals for different height values to indicate that the model assumptions are not reasonable.

But consider what happens when the relationship between insertion depth and weight is examined. A scatterplot of insertion depth and weight (kg) is shown in Figure 13.17(a), and a standardized residual plot in Figure 13.17(b). While some curvature is evident in the original scatterplot, it is even more clearly visible in the standardized residual plot. A careful inspection of these plots suggests that along with curvature, the residuals may be more variable at larger weights. When plots have this curved appearance and increasing variability in the residuals, the linear regression model is not appropriate.

Figure 13.17
(a) Scatterplot for insertion depth vs. weight data of Example 13.7; (b) standardized residual plot.



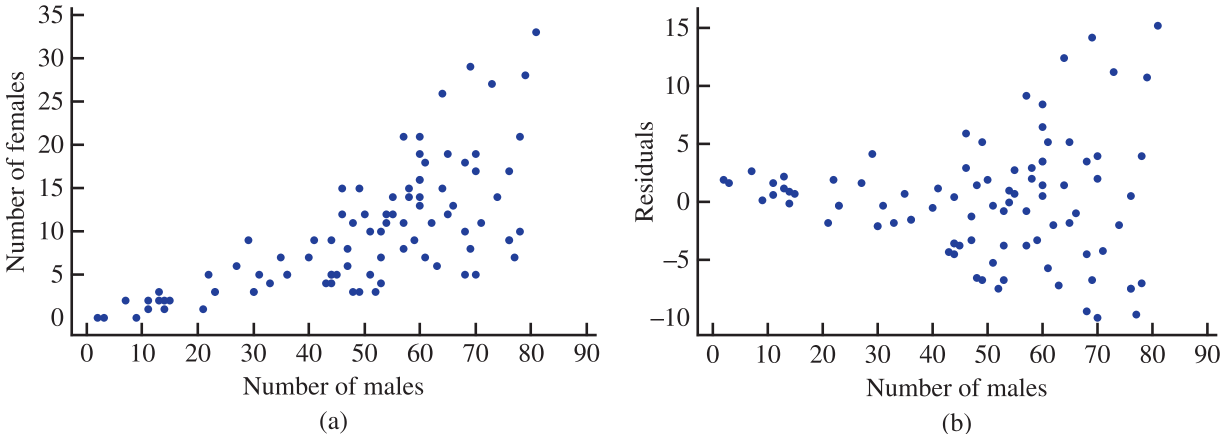
(a) (b)

Example 13.8 Looking for Love in All the Right... Trees?

Understand the context ➤

Treefrogs' search for mating partners was the examined in the article, **"The Cause of Correlations Between Nightly Numbers of Male and Female Barking Treefrogs (*Hyla gratiosa*) Attending Choruses" (*Behavioral Ecology* [2002]: 274–281)**. A "lek," in the world of animal behavior, is a cluster of males gathered in a relatively small area to exhibit courtship displays. The "female preference" hypothesis asserts that females will prefer larger leks over smaller leks, presumably because there are more males to choose from. The scatterplot and residual plot in Figure 13.18 show the relationship between the number of females and

Figure 13.18
(a) Scatterplot for treefrog data of Example 13.8; (b) residual plot.



(a) (b)

Copyright 2025 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s). Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

the number of males in observed leks of barking treefrogs. The unequal vertical variability is noticeable in the scatterplot and is even more evident in the residual plot. This indicates that the assumptions of the linear regression model are not reasonable in this situation.

Example 13.9 The Business of Baseball

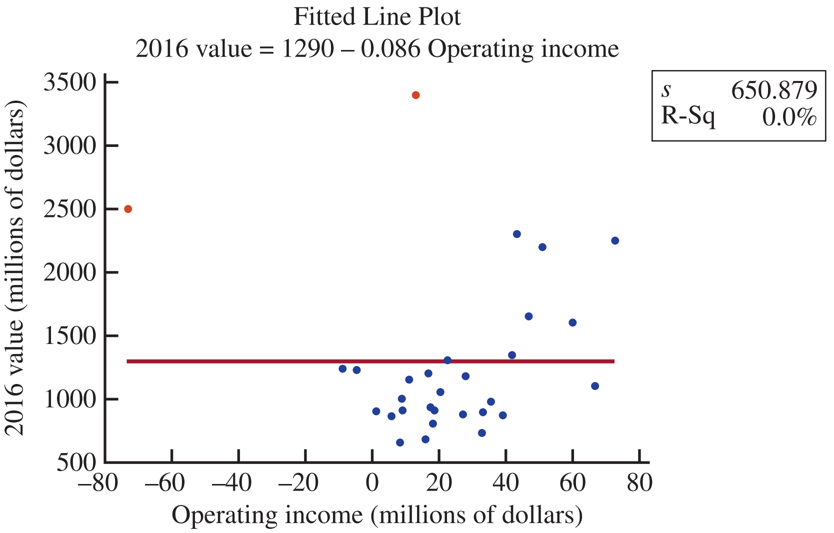
● The article **"The Business of Baseball" (forbes.com)** ranked the 30 major league baseball teams based on their 2016 value (in millions of dollars). Also included in the article are data on annual operating income (in millions of dollars). A positive value for operating income indicates a profit for the year and a negative operating income represents a loss for the year.

Team	2016Value (millions of dollars)	Operating Income (millions of dollars)
New York Yankees	3,400	13.0
Los Angeles Dodgers	2,500	−73.2
Boston Red Sox	2,300	43.2
San Francisco Giants	2,250	72.6
Chicago Cubs	2,200	50.8
New York Mets	1,650	46.8
St. Louis Cardinals	1,600	59.8
Los Angeles Angels of Anaheim	1,340	41.7
Washington Nationals	1,300	22.5
Philadelphia Phillies	1,235	−8.9
Texas Rangers	1,225	−4.7
Seattle Mariners	1,200	16.8
Atlanta Braves	1,175	27.8
Detroit Tigers	1,150	11.0
Houston Astros	1,100	66.6
Chicago White Sox	1,050	20.2
Baltimore Orioles	1,000	8.8
Pittsburgh Pirates	975	35.3
Arizona Diamondbacks	925	17.4
Minnesota Twins	910	18.5
Cincinnati Reds	905	9.0
Toronto Blue Jays	900	1.2
San Diego Padres	890	32.9
Milwaukee Brewers	875	27.0
Kansas City Royals	865	39.0
Colorado Rockies	860	5.5
Cleveland Indians	800	18.0
Oakland Athletics	725	32.7
Miami Marlins	675	15.8
Tampa Bay Rays	650	8.2

To investigate whether there is a relationship between y = 2016 value and x = Annual operating income, a simple linear regression model was fit. Figure 13.19 shows a scatterplot of the data and the least-squares regression line. Notice that there are two teams that stand out in the plot. One has an unusually low operating income (the Dodgers, with an operating income of −73.2 million dollars). The other team that stands out is a team with an unusually high 2016 value (the Yankees, with a 2016 value of 3400 million dollars [3.4 billion dollars]).

● **Data set available online**

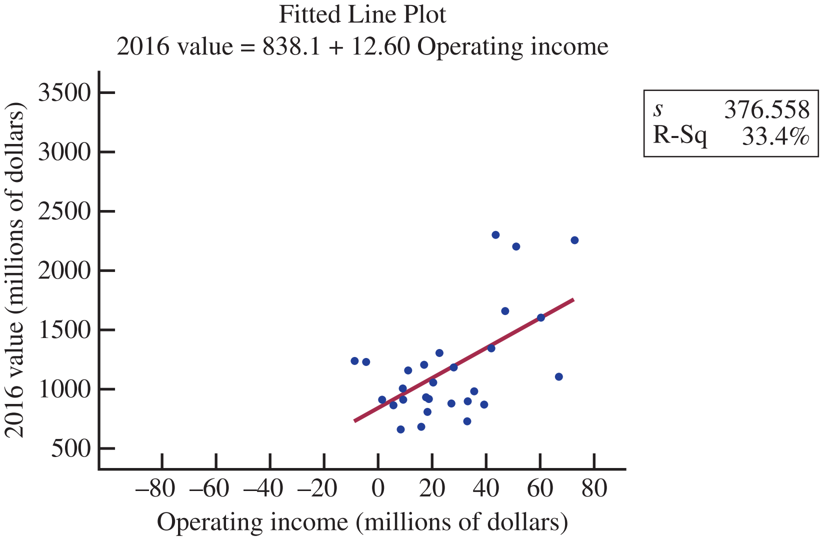
Figure 13.19
Scatterplot of 2016 value versus annual operating income for 30 major league baseball teams.



For this model, r^2 is approximately 0 and the null hypothesis that $\beta = 0$ is not rejected. This would lead you to think that there is not a useful linear relationship between 2016 value and operating income. But looking back at the scatterplot, you can see that if you were to ignore the two unusual teams, there does appear to be a positive linear relationship between 2016 value and operating income for the data set that consists of the remaining 28 teams. To investigate the potential influence of these two teams on the model, you can delete these two teams from the data set and then fit a regression model to the remaining data.

Figure 13.20 shows a scatterplot and the least-squares regression line for the 28 major league baseball teams that remain after the Dodgers and the Yankees are excluded from the data set. Notice that the slope of line has changed dramatically (from 0.086 to 12.60) and that the r^2 value is now 0.334. The model utility test confirms that there is a useful linear relationship between 2016 value and operating income for these 28 teams.

Figure 13.20
Scatterplot of 2016 value versus annual operating income for 28 major league baseball teams (Dodgers and Yankees excluded).



Occasionally, a residual plot or a standardized residual plot will display $\hat{y}$ plotted on the horizontal axis rather than x. Because $\hat{y}$ is just a linear function of x, using $\hat{y}$ rather than x changes the scale of the horizontal axis but does not change the pattern of the points in the plot. As a consequence, residual plots that use $\hat{y}$ on the horizontal axis can be interpreted in the same manner as residual plots that use x.

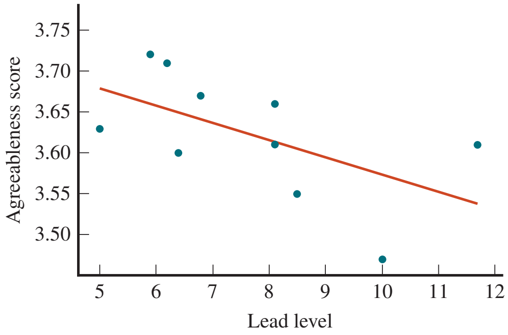
Copyright 2025 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s). Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

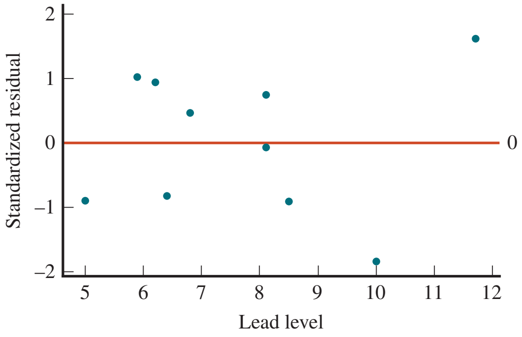
When the distribution of the random deviation e has heavier tails than the normal distribution, observations with large standardized residuals are not that unusual. Such observations can have a substantial effect on the equation of the estimated regression line when the least-squares approach is used. Statisticians have proposed a number of alternative methods—called **robust**, or **resistant**, methods—for fitting a line. These methods give less weight to outlying observations than does the least-squares method without deleting the outliers from the data set. The most widely used robust procedures require a substantial amount of computation, so an appropriate computer program is necessary.

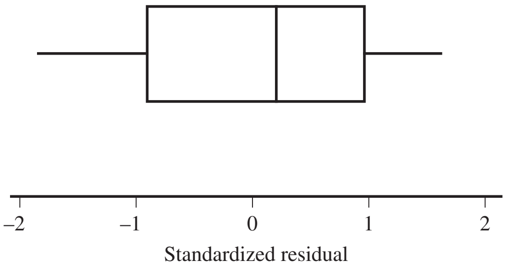
Exercises 13.31–13.36

● **Data set available online**

13.31 The following graphs are based on data from an investigation into the relationship between childhood lead exposure and personality development (**"Lead Exposure During Childhood Has Long-Lasting Effects," *Economist*, July 19, 2021**). Researchers used data from self-assessment personality questionnaires to measure personality traits, including agreeableness and extraversion. These personality trait measures were then paired with information on atmospheric-lead pollution based on the postal code of a participant's childhood home. The scatterplot, standardized residual plot, and a boxplot of the standardized residuals are shown here. Does it appear that the assumptions of the simple linear regression model are plausible? Explain your reasoning.





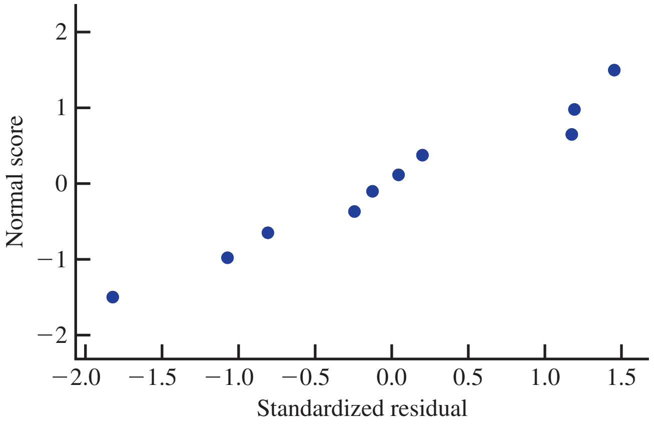


13.32 ● Exercise 13.23 gave data on x = Nerve firing frequency and y = Pleasantness rating when nerves were stimulated by a light brushing stroke on the forearm. The x values and the corresponding residuals from a simple linear regression are as follows:

Firing Frequency, x	Standardized Residual
23	−1.83
24	0.04
22	1.45
25	0.20
27	−1.07
28	1.19
34	−0.24
33	−0.13
36	−0.81
34	1.17

a. Construct a standardized residual plot. Does the plot exhibit any unusual features?

b. A normal probability plot of the standardized residuals follows. Based on this plot, is it reasonable to assume that the error distribution is approximately normal? Explain.



13.33 ● The paper **"Jumping in Arboreal Salamanders: A Possible Tradeoff Between Takeoff Velocity and In-Air Posture" (*Zoology* [2020]: 125724)** investigated jumping behavior of 10 species of salamanders and reported the following data on x = Bend angle

Copyright 2025 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s). Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

(the bend [in degrees] in the salamander body, with larger number reflecting more extreme bending) and y = Jump takeoff velocity (in meters per second). The accompanying table gives the observed x and y values and the corresponding standardized residuals for the estimated regression line.

Bend Angle	Takeoff Velocity	Standardized Residuals
37	0.6	0.13
55	0.7	−1.38
44	0.7	0.30
50	0.7	−0.63
70	0.9	−1.30
71	1.0	−0.10
67	1.1	1.87
48	0.7	−0.33
62	0.9	0.04
53	0.9	1.40

a. Construct a standardized residual plot. Are there any unusually large residuals? Do you think that there are any influential observations?

b. Is there any pattern in the standardized residual plot that would indicate that the simple liner regression model is not appropriate?

13.34 ● The **"2017 Chicago Community Area Economic Hardship Index" (Great Cities Institute, 2017)** includes data on various economic indicators for the different community areas that make up the city of Chicago. Data on x = Unemployment rate and y = Percent of housing units with more than one person per room for a random sample of 11 community areas are given in the accompanying table.

Unemployment Rate	Percent of Housing Units with More Than One Person per Room
1.6	6.7
6.4	11.8
8.2	9.2
6.1	7.0
3.6	10.1
2.0	5.1
9.1	12.0
5.5	12.0
1.1	3.1
10.9	12.7
13.6	8.7

a. Fit the simple linear regression model that would allow prediction of a community area's percent of housing units with more than one person per room based on its unemployment rate.

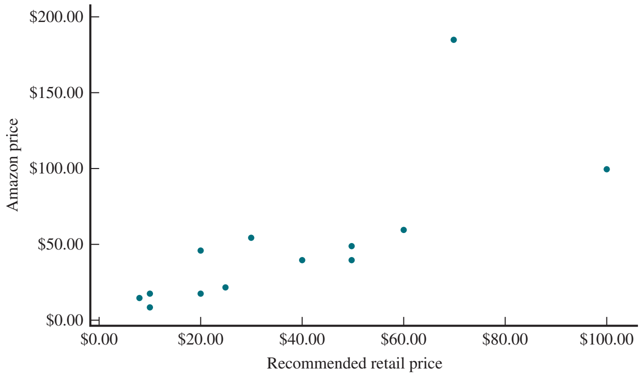
b. Calculate the standardized residuals (or just the residuals if you don't have access to a computer program that gives standardized residuals) and make a residual plot to determine whether there are any outliers.

c. The data point with the largest residual is for the Lower West Side community area. Delete this data point and refit the regression. Did deletion of this point result in a large change in the equation of the estimated regression line?

d. For the regression line of Part (c), interpret the estimated slope and, if appropriate, the estimated intercept.

e. For the data set with the Lower West Side community area deleted, do you think that the assumptions of the simple linear regression model are reasonable? Give statistical evidence for your answer.

13.35 The accompanying scatterplot is based on a random sample from a database of information on LEGO sets manufactured between 2018 and 2020 that was provided as part of the supplemental materials accompanying the article **"Building a Multiple Linear Regression Model with LEGO Brick Data" (*Journal of Statistics and Data Science Education* [2021]: 297–303).** The variables used to construct the plot were x = Retail price (in dollars) recommended by LEGO and y = Price (in dollars) of the LEGO set on Amazon. Discuss the effect that the observation (69.99, 184.95) might have on the equation that the estimated regression line. If this point were omitted, what do you think would happen to the slope of the estimated regression line compared to the slope when this point is included?



13.36 ● The random sample referenced in the previous exercise also included data on x = number of pieces in a LEGO set and y = retail price (in dollars) recommended by LEGO. The accompanying table gives the data on x and y. The standardized residuals for the estimated regression model are also given.

Number of Pieces	Recommended Retail Price	Standardized Residual
26	$19.99	0.3
55	$9.99	−0.7
73	$59.99	2.6
97	$19.99	−0.6
34	$19.99	0.2
7	$6.99	−0.4
40	$9.99	−0.5
15	$19.99	0.4
12	$9.99	−0.3
88	$9.99	−1.2

a. What assumptions are required for the simple linear regression model to be appropriate?

b. Construct a normal probability plot of the standardized residuals. Does the assumption that the random deviation distribution is normal appear to be reasonable? Explain.

c. Construct a standardized residual plot. Are there any unusually large residuals?

d. Is there anything about the standardized residual plot that would cause you to question the use of the simple linear regression model to describe the relationship between x and y?

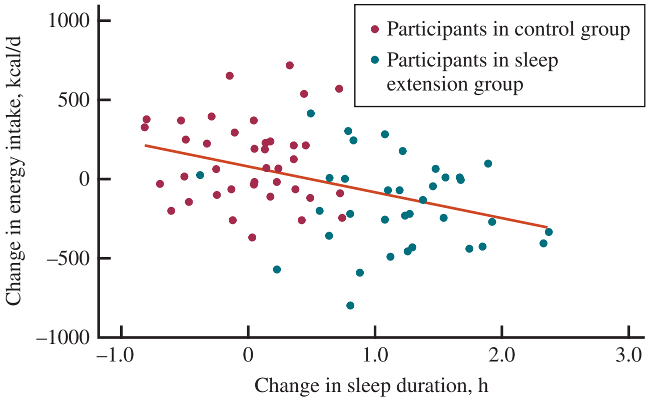
Chapter Activities

Activity 13.1 Sleep and Energy Intake

The purpose of this activity is to provide an opportunity to develop your multivariable thinking by taking a third variable into account when exploring linear relationships.

Background: The article **"Effect of Sleep Extension on Objectively Assessed Energy Intake Among Adults with Overweight in Real-Life Settings" (*JAMA Internal Medicine*, February 7, 2022)** describes a study of the relationship between sleep and energy intake in overweight individuals. In this study, 80 participants were randomly assigned to two groups—a control group and a sleep extension group. Those in the sleep extension group participated in an individualized counseling session intended to help them extend their sleep time to 8.5 hours per night. Those in the control group did not receive this counseling. Energy intake and sleep duration for the 2-week period prior to any counseling sessions and for the 2-week period after counseling sessions occurred were measured. These measurements were used to calculate the changes used to construct the accompanying graph.

1. The line drawn in the scatterplot is the estimated regression line for the complete data set that included participants in the control group as well as participants in the sleep extension group. If the estimated regression line had been fit to only those in the control group, how do you think the line would be different?
2. If estimated regression lines were calculated for each group separately (the control group and the sleep extension group), how do you think the lines would differ? Do you think that the slopes of the two lines would be similar?
3. Write a paragraph discussing what you learn from the given graph about the relationship between change in sleep duration and change in energy intake for the two groups.


Change from baseline in energy intake vs change from baseline in sleep duration

Copyright 2025 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s). Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

Summary Key Concepts and Formulas

Term or Formula	Comment
Simple linear regression model, $y = \alpha + \beta x + e$	This model assumes that there is a line with slope β and y intercept α, called the population regression line, such that an observation deviates from the line by a random amount e. The random deviation is assumed to have a normal distribution with mean zero and standard deviation σ, and random deviations for different observations are assumed to be independent of one another.
Estimated regression line, $\hat{y} = a + bx$	The least-squares line introduced in Chapter 5.
$s_e = \sqrt{\dfrac{\text{SSResid}}{n-2}}$	The point estimate of the standard deviation σ, with associated degrees of freedom $(n - 2)$.
$s_b = \dfrac{s_e}{\sqrt{\sum(x - \bar{x})^2}}$	The estimated standard deviation of the statistic b.
$b \pm (t \text{ critical value})s_b$	A confidence interval for the slope β of the population regression line, where the t critical value is based on $(n - 2)$ degrees of freedom.
$t = \dfrac{b - \text{hypothesized value}}{s_b}$	The test statistic for testing hypotheses about β. The test is based on $(n - 2)$ degrees of freedom.
Model utility test, with test statistic $t = \dfrac{b}{s_b}$	A test of H_0: $\beta = 0$, which asserts that there is no useful linear relationship between x and y, versus H_a: $\beta \neq 0$, the claim that there is a useful linear relationship.
Residual analysis	Methods based on the residuals or standardized residuals for checking the assumptions of a regression model.
Standardized residual	A residual divided by its standard deviation.
Standardized residual plot	A plot of the (x, standardized residual) pairs. A pattern in this plot suggests that the simple linear regression model may not be appropriate.

Technology Notes

Test for Slope of Regression Line

TI-83/84

1. Enter the data for the independent variable into **L1** (In order to access lists press the **STAT** key, highlight the option called **Edit…** then press **ENTER**)
2. Enter the data for the dependent variable into **L2**
3. Press **STAT**
4. Highlight **TESTS**
5. Highlight **LinRegTTest…** and press **ENTER**
6. Next to β **&** ρ select the appropriate alternative hypothesis
7. Highlight **Calculate** and press **ENTER**

TI-Nspire

1. Enter the data into two separate data lists (In order to access data lists select the spreadsheet option and press **enter**) **Note:** Be sure to title the lists by selecting the top row of the column and typing a title.
2. Press the **menu** key and select **4:Statistics** then **4:Stats Tests** then **A:Linear Reg t Test…** and press **enter**
3. In the box next to **X List** choose the list title where the independent data are stored from the drop-down menu
4. In the box next to **Y List** choose the list title where the dependent data are stored from the drop-down menu
5. In the box next to **Alternate Hyp** choose the appropriate alternative hypothesis from the drop-down menu
6. Press **enter**

JMP

1. Input the data for the dependent variable into the first column
2. Input the data for the independent variable into the second column
3. Click **Analyze** and select **Fit Y by X**
4. Select the dependent variable (Y) from the box under **Select Columns** and click on **Y, Response**
5. Select the independent variable (X) from the box under **Select Columns** and click on **X, Factor**
6. Click **OK**
7. Click the red arrow next to **Bivariate Fit of…** and select **Fit Line**

Minitab

1. Input the data for the dependent variable into the first column
2. Input the data for the independent variable into the second column
3. Select **Stat** then **Regression** then **Regression** then **Fit Regression Model…**

4. Click in the box labeled **Responses:** then highlight the name of the column containing the dependent variable and click **Select**
5. Click in the box labeled **Continuous predictors:** then highlight the name of the column containing the independent variable and click **Select**
6. Click **OK**

Note: Scroll up in the Session window to view the *t*-test results for the regression analysis.

SPSS

1. Input the data for the dependent variable into one column
2. Input the data for the independent variable into a second column
3. Click **Analyze** then click **Regression** then click **Linear…**
4. Select the name of the dependent variable and click the arrow to move the variable to the box under **Dependent:**
5. Select the name of the independent variable and click the arrow to move the variable to the box under **Independent(s):**
6. Click **OK**

Note: The *p*-value for the regression test can be found in the **Coefficients** table in the row with the independent variable name.

Excel

1. Input the data for the dependent variable into the first column
2. Input the data for the independent variable into the second column
3. Click on the **Data** ribbon
4. Click **Data Analysis** in the **Analysis** group

Note: If you do not see **Data Analysis** listed on the Ribbon, see the Technology Notes for Chapter 2 for instructions on installing this add-on.

5. Select **Regression** from the dialog box and click **OK**
6. Click in the box next to **Input Y Range:** and select the column of data containing the dependent variable
7. Click in the box next to **Input X Range:** and select the column of data containing the independent variable (if you have used and selected column titles for BOTH variables, select the check box next to **Labels**)
8. Click **OK**

Note: The test statistic and *p*-value for the regression test for the slope can be found in the third table of output. These values are listed in the row titled with the independent variable name and the columns entitled *t Stat* and *P-value*.

Cumulative Review Exercises CR13.1–CR13.13

● Data set available online

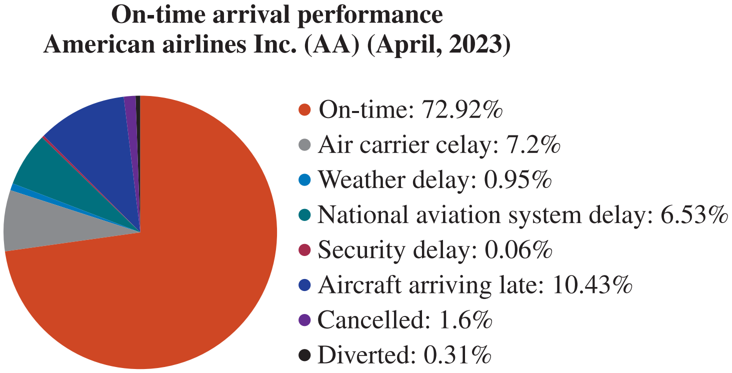
CR13.1 The MIND diet is a diet that emphasizes fresh fruit and berries, leafy green vegetables, and legumes such as beans, peas, and lentils. This diet was originally developed to help ward off cognitive decline in older adults. Researchers at the University of Illinois wondered if this diet could lead to improvement in children's cognitive development **("MIND Diet Linked with Better Focus in School-Aged Children," eurekalert.org, July 23, 2023)**. They studied 85 children age 7 to 11. Each child completed a 7-day diet record. They also completed a task that required spatial attention and their reaction time and accuracy on the task was recorded.

a. Was this study an observational study or an experiment?

b. The summary of the research noted that there was a positive relationship between the MIND diet score (a measure of the extent to which they followed the MIND diet) and accuracy on the task. The article states "Study participants who better adhered to the MIND diet performed better on the task." Is it reasonable to conclude that adherence to a MIND diet is the cause of better performance?

CR13.2 The Bureau of Transportation Statistics reports data on the on-time arrival performance for different airlines. Data for the month of April 2023 were used to produce the accompanying pie charts for three airlines—American, United, and Spirit (**transtats.bts.gov**). Write a few sentences commenting on the similarities and differences in on-time arrival performance for these three airlines.

Figure for Exercise CR13.2

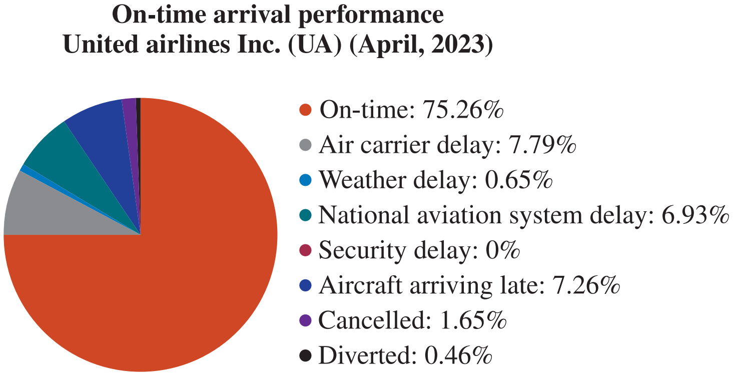


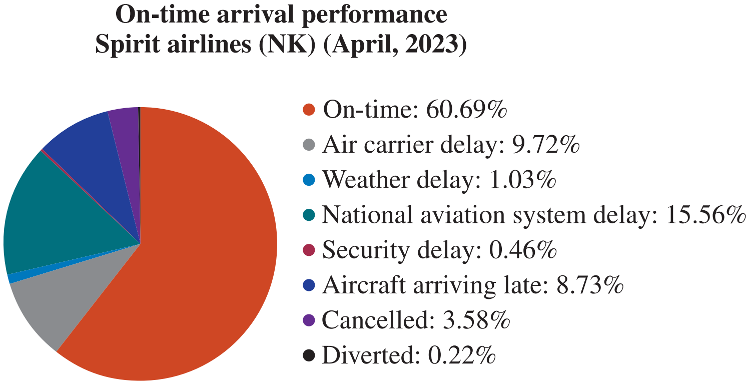
On-time arrival performance
American airlines Inc. (AA) (April, 2023)

(continued)

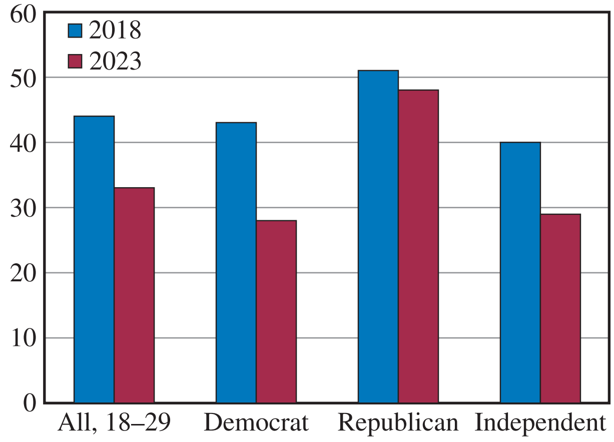
Copyright 2025 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s). Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

Figure for Exercise CR13.2 continued





CR13.3 Each year, researchers at Harvard University conduct a survey of a representative sample of approximately 2000 young adults age 18 to 29. In both 2018 and 2023, one of the questions on the survey was "How often do you trust the Supreme Court to do the right thing?" Survey participants were also classified according to their political affiliation. The accompanying graph is similar to one that appeared in the survey report **("Harvard Youth Poll," Spring 2023)**, and shows the percentages who responded all of the time or most of the time to the question about trusting the Supreme Court to do the right thing.



a. Based on the graph, describe what you see in terms of change over time (2018 compared to 2023).

b. Based on the graph, describe what you see in terms of both political affiliation and change over time.

CR13.4 The article **"Odds Are, It's Wrong" (*Science News*, March 27, 2010)** poses the following scenario:

> Suppose that a test for steroid use among baseball players is 95% accurate—that is, it correctly identifies actual steroid users 95% of the time, and misidentifies non-users as users 5 percent of the time. … Now suppose, based on previous testing, that experts have established that about 5 percent of professional baseball players use steroids.

Answer the following questions for this scenario.

a. If 400 professional baseball players are selected at random, how many would be expected to be steroid users and how many would be expected to be nonusers?

b. How many of the steroid users would be expected to test positive for steroid use?

c. How many of the players who do not use steroids would be expected to test positive for steroid use (a false positive)?

d. Use the answers to Parts (b) and (c) to estimate the proportion of those who test positive for steroid use who actually do use steroids.

Copyright 2025 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s). Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

e. Write a few sentences explaining why, in this scenario, the proportion of those who test positive for steroid use who actually use steroids is not 0.95.

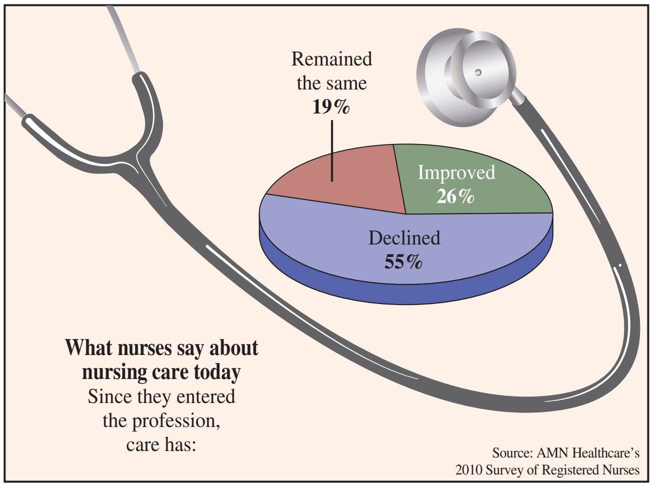
CR13.5 The article **"Nearly Half of Young Adults are Living at Home with Their Parents, Spending More on Luxury Goods: Report" (foxbusiness.com, December 15, 2022)** stated that 48% of the people in a sample of young adults age 18 to 29 said that they were living at home with their parents.

a. What are two things you would want to know about this sample before using data from the sample to draw conclusions about the proportion of young adults who live at home with their parents?

b. Suppose that the sample size was 1200 and that the other assumptions for inference are met. Do the sample data provide convincing evidence that the proportion of young adults who are living at home with their parents is less than 0.50? Test the relevant hypotheses using $\alpha = 0.05$.

CR13.6 The article referenced in the previous exercise also reported that in a sample of 1200 adults age 30 to 41, about 1 in 4 were living at home with their parents. Suppose that the actual sample proportion was 0.25, and that the sample was representative of adults in this age group. Construct and interpret a 95% confidence interval for the difference in the proportion that are living at home with their parents for the two age groups.

CR13.7 The accompanying graphical display is similar to one that appeared in ***USA TODAY* (February 19, 2010)**. It is meant to be a pie chart, but an oval rather than a circle is used to represent the whole pie. Does this graph do a good job of representing the proportion falling into each of the three response categories? Explain why or why not.



What nurses say about nursing care today
Since they entered the profession, care has:

Source: AMN Healthcare's 2010 Survey of Registered Nurses

CR13.8 To raise revenues, many airlines now charge fees to check luggage. Suppose that the number of checked bags was recorded for each person in a random sample of 100 airline passengers selected before fees were imposed and also for each person in a random sample of 100 airline passengers selected after fees were imposed, resulting in the accompanying data. Do the data provide convincing evidence that the proportions in each of the number of checked bags categories are not all the same before and after fees were imposed? Test the appropriate hypotheses using a significance level of 0.05.

	Number of Checked Bags		
	0	**1**	**2 or more**
Before fees	7	70	23
After fees	22	64	14

CR13.9 ● Consider the following data on y = Number of movies watched on a streaming service and x = Number of months the user has subscribed to the streaming service for a sample of 15 subscribers.

x	y
23	48
35	75
2	8
28	58
5	11
32	72
23	45
10	13
4	6
26	47
1	3
8	12
13	26
9	13
5	14

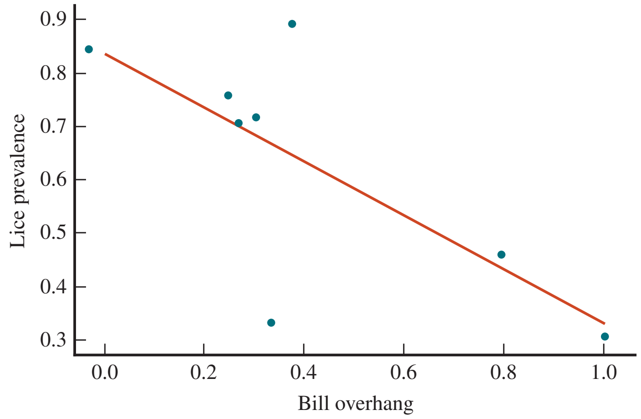
a. Construct a scatterplot of the data. Does the relationship between x and y look approximately linear?

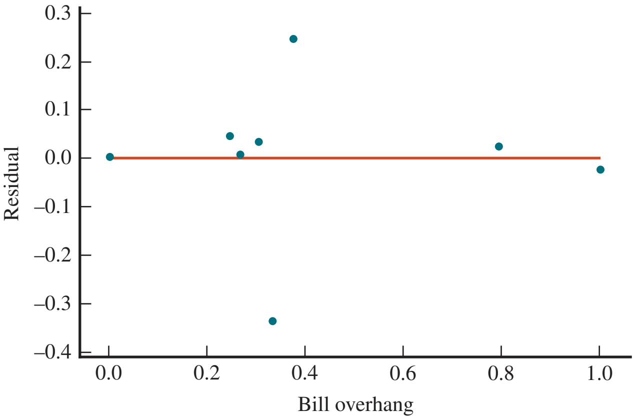
b. What is the equation of the estimated regression line?

c. Are the assumptions of the simple linear regression model reasonable? Justify the answer using appropriate graphs.

d. Is the simple linear regression model useful for describing the relationship between x and y? Test the relevant hypotheses using a significance level of 0.05.

CR13.10 Researchers in Hawaii have documented a large increase in the prevalence of a bird parasite known as chewing lice **("Explosive Increase in Ectoparasites in Hawaiian Forest Birds," *Journal of Parasitology* [2008]: 1009–1021)**. Current data suggest that the prevalence of chewing lice may be less for bird species with a high degree of bill overhang. A species is said to have bill overhang when the upper bill extends downward in front of the end of the lower bill. The scatterplot on the next page shows the relationship between the prevalence of chewing lice and bill overhang for eight bird species in the Hawaiian Islands. A residual plot is also shown on the next page. Use these plots to identify any outliers or potentially influential observations. For each point you identify, assess its influence on the estimated slope of the regression line.





CR13.11 ● Each year, researchers at Harvard University conduct a survey of a representative sample of approximately 2000 young adults age 18 to 29. In both 2011 and 2022, survey participants were asked about whether they thought same-sex relationships were morally wrong. Possible responses were yes, no, and unsure. The accompanying table provides data consistent with summary quantities given in the survey report **("Harvard Youth Poll," Spring 2022)**.

	Response		
Year	**Yes**	**No**	**Unsure**
2011	313	525	362
2022	276	624	300

a. Is there convincing evidence that the response distributions are not the same for 2011 and 2022?

b. In the summary of findings from the survey, the authors state that the data indicates growing support for same-sex relationships. Is this conclusion consistent with the results of the hypothesis test in Part (a)? Explain.

CR13.12 The discharge of industrial wastewater into rivers affects water quality. To assess the effect of a particular power plant on water quality, 24 water specimens were taken 16 km upstream and 4 km downstream of the plant. Alkalinity (mg/L) was determined for each specimen, resulting in the summary quantities in the accompanying table. Do the data suggest that the actual mean alkalinity is higher downstream than upstream by more than 50 mg/L? Use a 0.05 significance level.

Location	***n***	**Mean**	**Standard Deviation**
Upstream	24	75.9	1.83
Downstream	24	183.6	1.70

CR13.13 ● It is hypothesized that when homing pigeons are disoriented in a certain manner, they will exhibit no preference for any direction of flight after takeoff. To test this, 120 pigeons are disoriented and released, and the direction of flight of each is recorded. The resulting data are given in the accompanying table.

Direction	**Frequency**
0° to $< 45°$	12
45° to $< 90°$	16
90° to $< 135°$	17
135° to $< 180°$	15
180° to $< 225°$	13
225° to $< 270°$	20
270° to $< 315°$	17
315° to $< 360°$	10

Use the goodness-of-fit test with significance level 0.10 to determine whether the data are consistent with this hypothesis.

Chapter 14

Multiple Regression Analysis

FREDERIC J. BROWN/AFP/Getty Images

The general objective of regression analysis is to model the relationship between a dependent variable *y* and one or more independent variables (also sometimes called predictor or explanatory variables). The simple linear regression model $y = \alpha + \beta x + e$, introduced in Chapter 13, relates *y* to a single independent variable *x*. In many situations, the relationship between *y* and any single independent variable is not strong, but knowing the values of several independent variables may considerably reduce uncertainty about the associated *y* value.

For example, some variability in house prices in a large city can be attributed to house size, but knowledge of size by itself would not usually allow a bank appraiser to accurately predict a home's value. Price is also determined to some extent by other variables, such as the age of the house, lot size, number of bedrooms and bathrooms, and distance from schools.

In this chapter, the simple linear regression model introduced in the previous chapter is extended to *multiple regression models*, which include at least two independent variables. Fortunately, many of the concepts developed in the context of simple linear regression carry over to multiple regression with little or no modification. However, the calculations required to fit a multiple regression model are *much* more time consuming than those for simple linear regression, so technology is an indispensable tool.

Learning Objectives

After completing this chapter, you will be able to:

- Assess the usefulness of a multiple regression model.
- Interpret the parameters of the multiple linear regression model in context.
- Use both numerical and categorical variables in a multiple regression model.
- Estimate the parameters in a multiple regression model.
- Explain the concept of interaction in a multiple regression setting.

Section 14.1 Multiple Regression Models

The relationship between a dependent variable y and two or more independent variables is deterministic if the value of y is completely determined, with no uncertainty, once values of the independent variables have been specified. For example, consider a school district in which teachers with no prior teaching experience and no college credits beyond a bachelor's degree start at an annual salary of $58,000. Suppose that for each year of teaching experience up to 20 years, a teacher receives an additional $800 per year and that each unit of postcollege coursework up to 75 units results in an extra $60 per year. Consider the following three variables:

y = Salary of a teacher who has at most 20 years of teaching experience and at most 75 postcollege units
x_1 = Number of years of teaching experience
x_2 = Number of postcollege units

Previously, x_1 and x_2 denoted the first two observations for a single variable x. In the usual notation for multiple regression, however, x_1 and x_2 represent two different variables.

For these variables, the value of y is entirely determined by values of x_1 and x_2 using the equation

$$y = 58{,}000 + 800x_1 + 60x_2$$

If $x_1 = 10$ and $x_2 = 30$ then

$$\begin{aligned} y &= 58{,}000 + 800(10) + 60(30) \\ &= 58{,}000 + 8000 + 1800 \\ &= 67{,}800 \end{aligned}$$

If two different teachers both have the same x_1 values and the same x_2 values, they will also have identical y values.

In practice, y is rarely deterministically related to predictors $x_1, \ldots, x_k$. A probabilistic model is more realistic in most situations. A probabilistic model results from adding a random deviation e to a deterministic function of the x_i's.

Definition

General additive multiple regression model: A model that relates a dependent variable y to k independent variables $x_1, x_2, \ldots, x_k$.

The model is specified by the model equation

$$y = \alpha + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k + e$$

Model assumptions for the multiple regression model are given in the following box:

The random deviation e is assumed to be normally distributed with mean value 0 and standard deviation σ for any particular values of $x_1, \ldots x_k$.

This implies that for fixed $x_1, x_2, \ldots, x_k$ values, y has a normal distribution with standard deviation σ and

$$\begin{pmatrix} \text{mean } y \text{ value for fixed} \\ x_1, \ldots, x_k \text{ values} \end{pmatrix} = \alpha + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k$$

The β_i's in the multiple regression model are called **population regression coefficients**. Each β_i can be interpreted as the mean change in y when the predictor x_i increases by 1 unit *and* the values of all the other predictors remain fixed.

Copyright 2025 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s). Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

$\alpha + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k$ is called the **population regression function**.

As in simple linear regression, if σ (the standard deviation of the random error distribution) is quite close to 0, any particular observed y will tend to be quite near its mean value. When σ is large, y observations may deviate substantially from their mean y values.

Example 14.1 Student Satisfaction with Online Courses

Understand the context ➤

What factors influence student satisfaction and perceived learning in online courses? To investigate, the authors of the paper **"Predicting Student Satisfaction and Perceived Learning Within Online Learning Environments" (*Distance Education* [2019]: 133–148)** collected data from a sample of 167 college students who took an online course. They considered the extent to which each of these students interacted with the course content and with the course instructor.

Consider the following variables

y = A measure of student satisfaction, with values ranging from 1 to 5
x_1 = A measure of student-content interaction, with values ranging from 1 to 5
x_2 = A measure of student-instructor interaction, with values ranging from 1 to 5

One possible population model might be

$$y = -0.4 + 0.8x_1 + 0.3x_2 + \varepsilon$$

The population regression function would then be

$$(\text{mean } y \text{ value for fixed } x_1, x_2) = -0.4 + 0.8x_1 + 0.3x_2$$

Interpret the results ➤

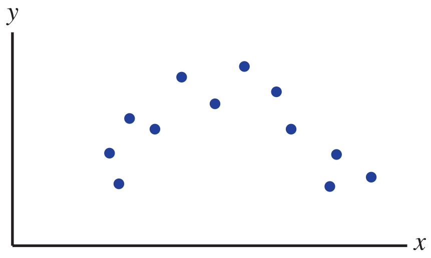
For students whose level of interaction with the course content was rated at 3.2 and whose level of engagement with the instructor was 3.8,

$$(\text{mean student satisfaction value}) = -0.4 + 0.8(3.2) + 0.3(3.8) = 3.30$$

A Special Case: Polynomial Regression

Consider again the case of a single independent variable x, and suppose that a scatterplot of the n sample (x, y) pairs has the appearance of Figure 14.1. The simple linear regression model is clearly not appropriate, but it does look as though a parabola (quadratic function) with equation $y = \alpha + \beta_1 x + \beta_2 x^2$ would provide a good fit to the data.

Figure 14.1
A scatterplot that suggests the appropriateness of a quadratic probabilistic model.



Just as the inclusion of the random deviation e in simple linear regression allowed an observation to deviate from the population regression line by a random amount, adding e to this quadratic function yields a probabilistic model in which an observation is allowed to fall above or below the parabola. The quadratic regression model equation is

$$y = \alpha + \beta_1 x + \beta_2 x^2 + e$$

The model equation can be rewritten using x_1 to denote x and x_2 to denote x^2. The model equation then becomes

$$y = \alpha + \beta_1 x_1 + \beta_2 x_2 + e$$

This is a special case of the general multiple regression model with $k = 2$. At first it may seem odd to allow one predictor variable to be a mathematical function of another predictor—here, $x_2 = (x_1)^2$. However, there is nothing in the general multiple regression model that prevents this. In the model $y = \alpha + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k + e$, it is possible to have predictors that are mathematical functions of other predictors.

For example, starting with the two independent variables x_1 and x_2, you could create a model with $k = 4$ predictors in which x_1 and x_2 themselves are the first two predictor variables and $x_3 = (x_1)^2$, $x_4 = x_1 x_2$. (The consequences of using a predictor such as x_4 will be explained later in this section.)

The general polynomial regression model begins with a single independent variable x and creates predictors $x_1 = x, x_2 = x^2, x_3 = x^3, \ldots, x_k = x^k$ for some specified value of k.

The kth-degree polynomial regression model

$$y = \alpha + \beta_1 x + \beta_2 x^2 + \cdots + \beta_k x^k + e$$

is a special case of the general multiple regression model with

$$x_1 = x \quad x_2 = x^2 \quad x_3 = x^3 \quad \ldots \quad x_k = x^k$$

The **population polynomial regression function** (the mean value of y for a fixed value of the predictor) is

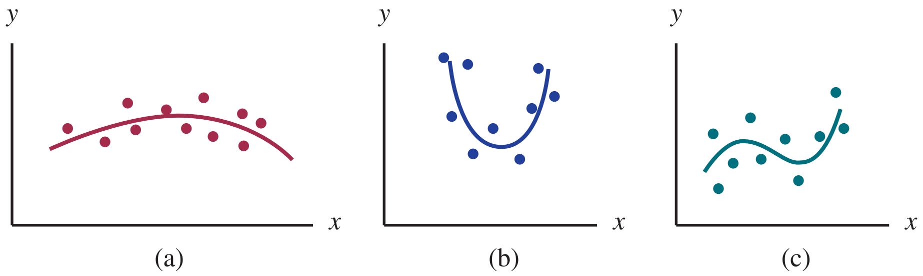
$$\alpha + \beta_1 x + \beta_2 x^2 + \cdots + \beta_k x^k$$

The most important special case other than simple linear regression ($k = 1$) is the **quadratic regression model**

$$y = \alpha + \beta_1 x + \beta_2 x^2 + e$$

This model replaces the line of mean values $\alpha + \beta x$ in simple linear regression with a parabolic curve of mean values $\alpha + \beta_1 x + \beta_2 x^2$. If $\beta_2 > 0$, the curve opens upward, and if $\beta_2 < 0$, the curve opens downward.

A less frequently encountered special case is cubic regression, in which $k = 3$. (See Figure 14.2.)



(a) (b) (c)

Figure 14.2 Polynomial regression: (a) quadratic regression model with $\beta_2 < 0$; (b) quadratic regression model with $\beta_2 > 0$; (c) cubic regression model with $\beta_3 > 0$.

Example 14.2 Increased Risk of Heart Attack

Understand the context ➤

Many researchers have examined factors that are believed to contribute to the risk of heart attack. The authors of the paper **"Obesity and the Risk of Myocardial Infarction in 27,000 Participants from 53 Countries: A Case-Control Study" (*The Lancet* [2005]: 1640–1649)** found that hip-to-waist ratio was a better predictor of heart attacks than was body mass index.

Copyright 2025 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s). Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

The dependent variable in this study was a measure of heart attack risk (y). Larger values of y indicate greater risk of heart attack. The independent variable was hip-to-waist ratio (x). A scatterplot exhibited a curved relationship. A model consistent with summary values given in the paper is

$$y = 1.023 + 0.024x + 0.060x^2 + e$$

The population regression function is

$$\begin{pmatrix}\text{mean value of}\\ \text{heart attack risk measure}\end{pmatrix} = 1.023 + 0.024x + 0.060x^2$$

Interpret the results ➤ For example, if $x = 1.3$

$$\begin{pmatrix}\text{mean value of}\\ \text{heart attack risk measure}\end{pmatrix} = 1.023 + 0.024(1.3) + 0.060(1.3)^2 = 1.16$$

If $\sigma = 0.25$, then $2\sigma = 0.50$. This implies that it is likely that the heart attack risk measure for a person with a hip-to-waist ratio of 1.3 would be between 0.66 and 1.66.

The interpretation of β_i previously given for the general multiple regression model is not appropriate for polynomial regression. This is because all predictors are functions of the single variable x, so $x_i = x^i$ cannot be increased by 1 unit without changing the values of all the other predictor variables as well. The interpretation of regression coefficients requires extra care when some predictor variables are mathematical functions of other variables.

Interaction Between Variables

Suppose that an industrial chemist is interested in the relationship between

y = Product yield from a certain chemical reaction

and two independent variables,

x_1 = Reaction temperature

and

x_2 = Pressure at which the reaction is carried out.

The chemist initially suggests that for temperature values between 80 and 110 in combination with pressure values ranging from 50 to 70, the relationship can be well described by the multiple regression model

$$y = 1200 + 15x_1 - 35x_2 + e$$

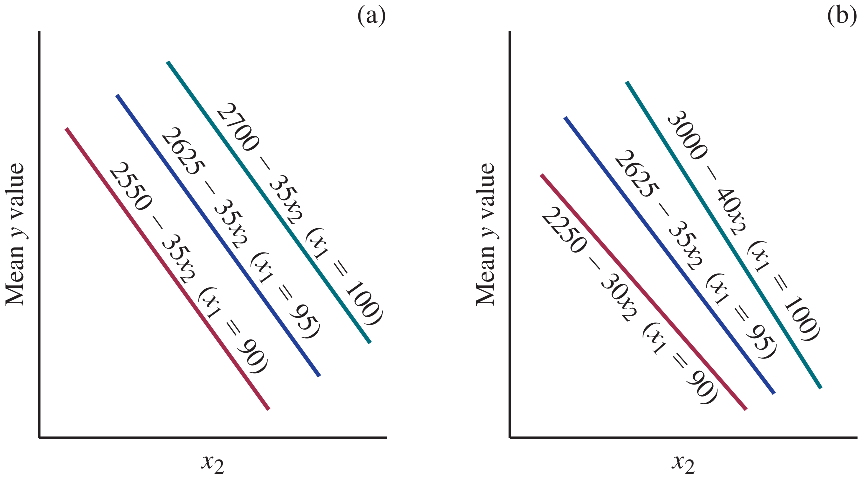
The regression function, which gives the mean y value for any specified values of x_1 and x_2, is then $1200 + 15x_1 - 35x_2$. Consider this mean y value for three different particular temperature values:

$x_1 = 90$: mean y value $= 1200 + 15(90) - 35x_2 = 2550 - 35x_2$
$x_1 = 95$: mean y value $= 2625 - 35x_2$
$x_1 = 100$: mean y value $= 2700 - 35x_2$

Graphs of these three mean value functions (each a function only of pressure x_2, because a particular temperature value has been specified) are shown in Figure 14.3(a). Each graph is a straight line, and the three lines are parallel, each one having slope -35. Because of this, the average change in yield when pressure x_2 is increased by 1 unit is -35, regardless of the fixed temperature value.

However, because theory suggests that the decline in average yield when pressure x_2 increases should be more rapid for a high temperature than for a low temperature, the chemist now has reason to doubt the appropriateness of the proposed model. Rather than

Figure 14.3
Graphs of mean y value for two different models:
(a) $1200 + 15x_1 - 35x_2$;
(b) $-4500 + 75x_1 + 60x_2 - x_1x_2$.



the lines being parallel, the line for a temperature of 100 should be steeper than the line for a temperature of 95, and that line should be steeper than the one for $x_1 = 90$.

A model that has this property includes a third independent variable, $x_3 = x_1x_2$, in addition to x_1 and x_2 separately. One such model is

$$y = -4500 + 75x_1 + 60x_2 - x_1x_2 + e$$

which has regression function $-4500 + 75x_1 + 60x_2 - x_1x_2$. Then

$$\begin{aligned}(\text{mean } y \text{ when } x_1 = 100) &= -4500 + 75(100) + 60x_2 - 100x_2\\ &= 3000 - 40x_2\end{aligned}$$

whereas

$$\begin{aligned}(\text{mean } y \text{ when } x_1 = 95) &= 2625 - 35x_2\\ (\text{mean } y \text{ when } x_1 = 90) &= 2250 - 30x_2\end{aligned}$$

These functions are graphed in Figure 14.3(b), where it is clear that the three slopes are different. In fact, each different value of x_1 yields a different slope, so the average change in yield associated with a 1-unit increase in x_2 depends on the value of x_1. When this is the case, the two variables are said to *interact*.

Definition

Interaction: If the change in the mean y value associated with a 1-unit increase in one independent variable depends on the value of a second independent variable, there is **interaction** between these two independent variables.

When the variables are denoted by x_1 and x_2, interaction can be modeled by including x_1x_2, the product of the variables that interact, as an independent variable in the model.

The general equation for a multiple regression model based on two independent variables x_1 and x_2 that also includes an interaction term is

$$y = \alpha + \beta_1x_1 + \beta_2x_2 + \beta_3x_1x_2 + e$$

When x_1 and x_2 do interact, this model usually gives a much better fit to the resulting sample data—and thus explains more variability in y—than does the no interaction model. Failure to consider a model with interaction may lead to an incorrect conclusion that there is no strong relationship between y and a set of independent variables.

More than one interaction term can be included in the model when there are more than two independent variables. For example, if there are three independent variables x_1, x_2, and x_3, one possible model is

$$y = \alpha + \beta_1x_1 + \beta_2x_2 + \beta_3x_3 + \beta_4x_4 + \beta_5x_5 + \beta_6x_6 + e$$

where

$$x_4 = x_1x_2 \qquad x_5 = x_1x_3 \qquad x_6 = x_2x_3$$

You could even include a three-way interaction variable $x_7 = x_1x_2x_3$ (the product of all three independent variables), although this is not done very often in practice.

In applied work, quadratic terms, such as x_1^2 and x_2^2 are often included to model a curved relationship between y and one or more independent variables. A frequently used model involving just two independent variables x_1 and x_2 but $k = 5$ predictors is the *full quadratic* or **complete second-order model**

$$y = \alpha + \beta_1x_1 + \beta_2x_2 + \beta_3x_1x_2 + \beta_4x_1^2 + \beta_5x_2^2 + e$$

This model replaces the straight lines of Figure 14.3 with parabolas (each one is the graph of the regression function for different values of x_2 when x_1 has a fixed value).

With four independent variables, a model containing four quadratic terms and six two-way interaction terms is possible. Clearly, with just a few independent variables, there are a great many different multiple regression models. Section 14.5 (online) includes a brief discussion of methods for selecting one model from a number of competing models.

When developing a multiple regression model, scatterplots of y with each potential predictor can be informative. This is illustrated in Example 14.3, which considers a model that includes a term that is a function of one of the independent variables and also an interaction term.

Example 14.3 Wind Chill Factor

Understand the context ➤

● The wind chill index, often included in winter weather reports, combines information on air temperature and wind speed to describe how cold it really feels. In 2001, the National Weather Service announced that it would begin using a new wind chill formula beginning in the fall of that year. The following table gives the wind chill index for various combinations of air temperature and wind speed.

	Temperature (°F)														
Wind Speed (mph)	**35**	**30**	**25**	**20**	**15**	**10**	**5**	**0**	**−5**	**−10**	**−15**	**−20**	**−25**	**−30**	**−35**
5	31	25	19	13	7	1	−5	−11	−16	−22	−28	−34	−40	−46	−52
10	27	21	15	9	3	−4	−10	−16	−22	−28	−35	−41	−47	−53	−59
15	25	19	13	6	0	−7	−13	−19	−26	−32	−39	−45	−51	−58	−64
20	24	17	11	4	−2	−9	−15	−22	−29	−35	−42	−48	−55	−61	−68
25	23	16	9	3	−4	−11	−17	−24	−31	−37	−44	−51	−58	−64	−71
30	22	15	8	1	−5	−12	−19	−26	−33	−39	−46	−53	−60	−67	−73
35	21	14	7	0	−7	−14	−21	−27	−34	−41	−48	−55	−62	−69	−76
40	20	13	6	−1	−8	−15	−22	−29	−36	−43	−50	−57	−64	−71	−78
45	19	12	5	−2	−9	−16	−23	−30	−37	−44	−51	−58	−65	−72	−79

Figure 14.4(a) shows a scatterplot of wind chill index versus air temperature with different wind speeds denoted by different colors. It appears that the wind chill index increases linearly with air temperature at each of the wind speeds, but the linear patterns for the different wind speeds are not quite parallel. This suggests that to model the relationship between

$$y = \text{Wind chill index}$$

and the two variables

$$x_1 = \text{Air temperature}$$

and

$$x_2 = \text{Wind speed,}$$

you should include both x_1 and an interaction term that involves x_2.

● **Data set available online**

Figure 14.4(b) shows a scatterplot of wind chill index versus wind speed with different temperatures denoted by different colors. This plot reveals that the relationship between wind chill index and wind speed is nonlinear at each of the different temperatures. Also, because the pattern is more markedly curved at some temperatures than at others, an interaction is suggested.

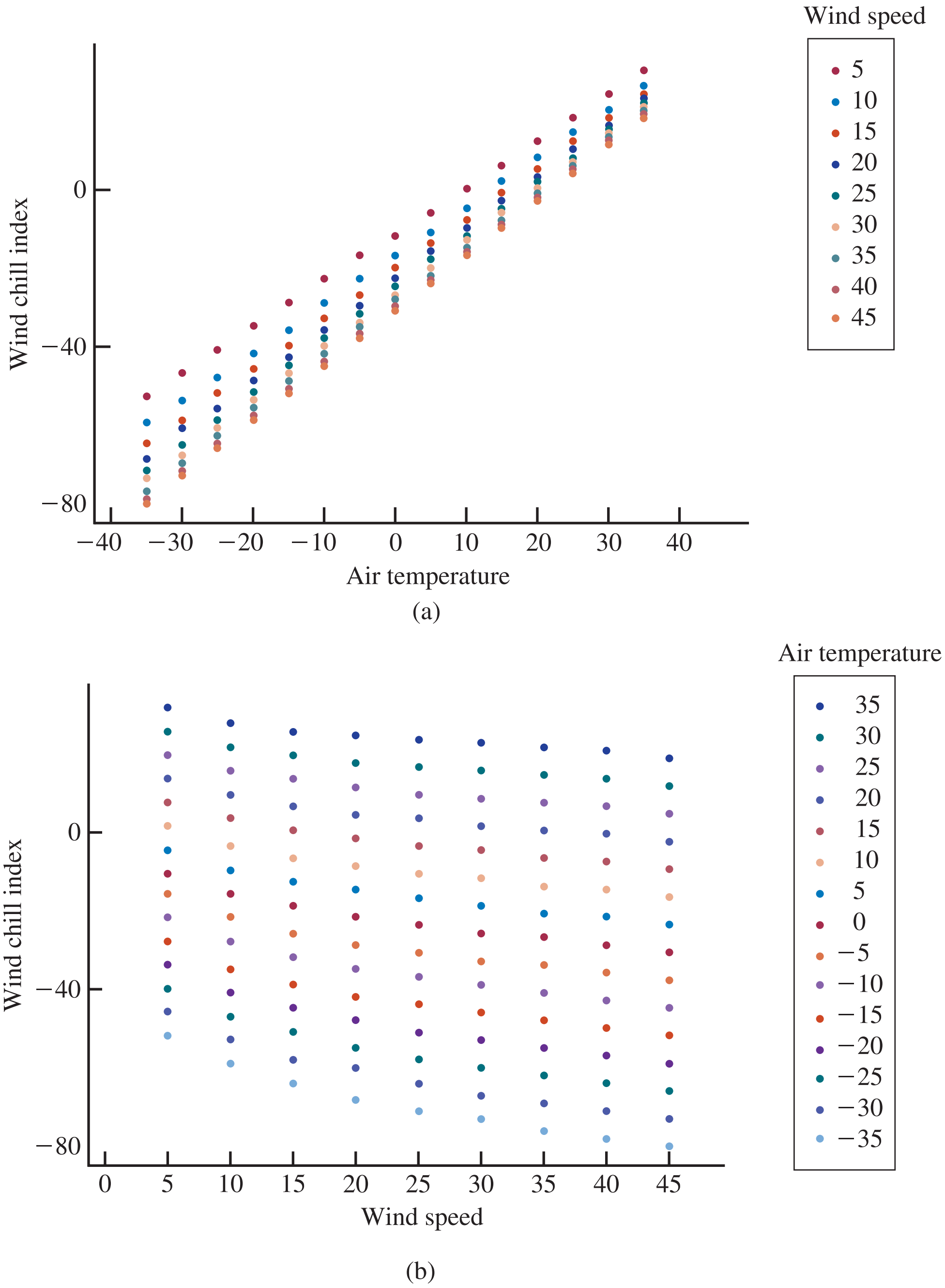


Figure 14.4
Scatterplots of wind chill index data of Example 14.3:
(a) wind chill index versus air temperature;
(b) wind chill index versus wind speed.

These observations are consistent with the new model used by the National Weather Service for relating wind chill index to air temperature and wind speed. The model used is

$$(\text{mean } y) = 35.74 + 0.621x_1 - 35.75(x_2') + 0.4275x_1x_2'$$

where

$$x_2' = x_2^{0.16}$$

Copyright 2025 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s). Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

which incorporates a variable that is a function of x_2 (to model the nonlinear relationship between wind chill index and wind speed) and an interaction term.

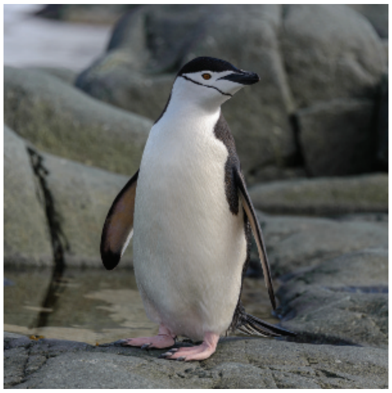
Qualitative Predictor Variables

Up to this point, you have only considered quantitative (numerical) predictor variables in a multiple regression model. Using a simple numerical coding, qualitative (categorical) variables can also be incorporated into a model. First, consider a dichotomous variable, one with just two possible categories, such as married or not married, vaccinated or not vaccinated, a house with or without a view, and so on. With a dichotomous variable, you associate a numerical variable x whose possible values are 0 and 1, where 0 is identified with one category (for example, married) and 1 is identified with the other possible category (for example, not married). This $0-1$ variable is often called an **indicator variable**.

Example 14.4 Predictors of Penguin Body Mass

Understand the context ➤

Body mass is one measure of the health of penguins. Researchers in Antarctica recorded data for a sample of 68 Chinstrap penguins (**"Palmer Archipelago Penguins Data in the palmerpenguins R Package - An Alternative to Anderson's Irises,"** ***The R Journal*****, 2022**). Three of the variables studied were body mass, flipper length, and sex. Because it is easier to carry a tape measure than a scale when visiting a penguin colony, it would be useful to have a way to estimate body mass using flipper length and sex as predictors. Note that while body mass and flipper length are numerical variables, sex is a qualitative predictor.



Alexey Seafarer/Shutterstock.com

Consider the following variables:

$$y = \text{Body mass (in grams)}$$

$$x_1 = \text{Sex} \begin{cases} 0 & \text{if female} \\ 1 & \text{if male} \end{cases}$$

$$x_2 = \text{Flipper length (in mm)}$$

One possible multiple regression model is

$$y = \alpha + \beta_1 x_1 + \beta_2 x_2 + e$$

Considering the mean y value first when $x_1 = 0$ and then when $x_1 = 1$ results in

$$\begin{array}{ll} (\text{mean body mass}) = \alpha + \beta_2 x_2 & \text{when } x_1 = 0 \text{ (female)} \\ (\text{mean body mass}) = \alpha + \beta_1 + \beta x_2 & \text{when } x_1 = 1 \text{ (male)} \end{array}$$

The coefficient β_1 is the difference in mean body mass between males and females when flipper length is held fixed.

A second possibility is a model with an interaction term:

$$y = \alpha + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1 x_2 + e$$

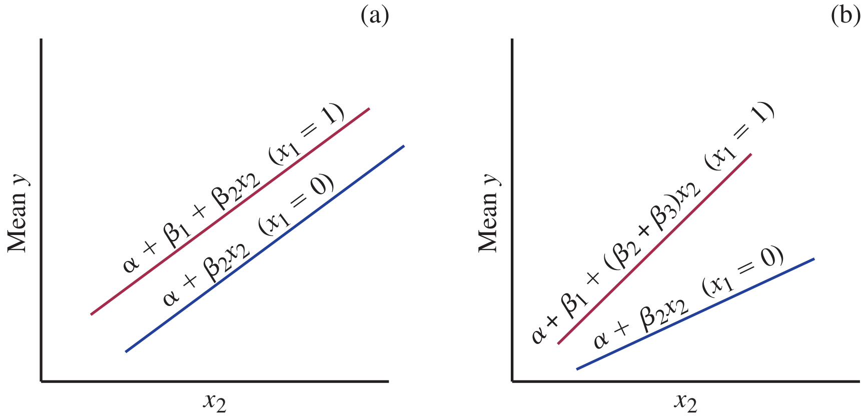
The regression function for this model is $\alpha + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3$ where $x_3 = x_1 x_2$. Now the two cases $x_1 = 0$ and $x_1 = 1$ result in

$$\begin{array}{ll} (\text{mean body mass}) = \alpha + \beta_2 x_2 & \text{when } x_1 = 0 \text{ (females)} \\ (\text{mean body mass}) = \alpha + \beta_1 + (\beta_2 + \beta_3) x_2 & \text{when } x_1 = 1 \text{ (males)} \end{array}$$

For each model, the graph of mean body mass, when regarded as a function of flipper length, is a line for females and a line for males (Figure 14.5).

In the no-interaction model, the coefficient of x_2 is β_2 both when $x_1 = 0$ and when $x_1 = 1$. This means that the two lines are parallel, although their intercepts are different (unless $\beta_1 = 0$). With interaction, the lines not only have different intercepts but also have different slopes (unless $\beta_3 = 0$). For the interaction model, the change in mean body mass when flipper length increases by 1 unit depends on sex. The two variables *flipper length* and *sex* interact.

Figure 14.5
Regression functions for models with one qualitative variable (x_1) and one quantitative variable (x_2):
(a) no interaction;
(b) interaction.



One might think that the way to incorporate a qualitative variable with three categories is to define a single numerical variable with coded values such as 0, 1, and 2 corresponding to the three categories. This is incorrect because it imposes an ordering on the categories that may not be appropriate. The correct approach to modeling a categorical variable with three categories is to define *two* different indicator variables, as illustrated in Example 14.5.

Example 14.5 Location, Location, Location

Understand the context ➤ One of the factors that has an effect on the price of a house is location. You might want to incorporate location, as well as numerical predictors such as size and age, into a multiple regression model for predicting house price.

Suppose that in a California beach community houses can be classified by location into three categories—ocean view and beachfront, ocean view but not beachfront, and no ocean view. Consider the following variables:

$$x_1 = \begin{cases} 1 & \text{if the house has an ocean view and is beachfront} \\ 0 & \text{otherwise} \end{cases}$$

$$x_2 = \begin{cases} 1 & \text{if the house has an ocean view but is not beachfront} \\ 0 & \text{otherwise} \end{cases}$$

x_3 = House size

x_4 = House age

This means that $x_1 = 1, x_2 = 0$ indicates a beachfront ocean-view house; $x_1 = 0, x_2 = 1$ indicates a house with an ocean view but not beachfront; and $x_1 = x_2 = 0$ indicates a house that does not have an ocean view. ($x_1 = x_2 = 1$ is not possible.) You could then consider a multiple regression model of the form

$$y = \alpha + \beta_1x_1 + \beta_2x_2 + \beta_3x_3 + \beta_4x_4 + e$$

This model allows individual adjustments to the predicted price for a house with no ocean view for the other two location categories. For example, β_1 is the amount that would be added to the predicted price for a home with no ocean view to adjust for a beachfront location (assuming that age and size were the same).

In general, incorporating a categorical variable with c possible categories into a regression model requires the use of $c - 1$ indicator variables. Even one such categorical variable can add many predictors to a model.

Exercises 14.1–14.15

14.1 **a.** Explain the difference between a deterministic and a probabilistic model.
b. Give an example of a dependent variable y and two or more independent variables that might be related to y in a deterministic way.
c. Give an example of a dependent variable y and two or more independent variables that might be related to y in a probabilistic way.

14.2 Example 14.1 describes a study investigating factors that influence student satisfaction and perceived learning in online courses (**"Predicting Student Satisfaction and Perceived Learning Within Online Learning Environments,"** ***Distance Education*** **[2019]: 133–148)**. Data were collected from a sample of 167 college students who took an online course. They used a multiple regression model to describe the relationship between the following variables:

y = A measure of the student's perception of how much they learned, with values ranging from 1 to 5
x_1 = A measure of student-content interaction, with values ranging from 1 to 5
x_2 = A measure of student-instructor interaction, with values ranging from 1 to 5
x_3 = A measure of student-student interaction, with values ranging from 1 to 5

a. The authors concluded that both student-content and student-instructor interaction were both important predictors of perceived learning but that student-student interaction was not. They also concluded that there was no significant interaction between any of the variables. What multiple regression model is consistent with this description?
b. The value of the coefficient of x_2 in the multiple regression function given in the paper is 0.31. Interpret this value.

14.3 A number of investigations have focused on the problem of assessing weights that can be lifted in a safe manner. The article **"Anthropometric, Muscle Strength, and Spinal Mobility Characteristics as Predictors in the Rating of Acceptable Loads in Parcel Sorting"** (***Ergonomics*** **[1992]: 1033–1044)** proposed using a regression model to relate the dependent variable

y = Individual's rating of acceptable weight (kg)
to $k = 3$ independent (predictor) variables:
x_1 = Extent of left lateral bending (cm)
x_2 = Dynamic hand grip endurance (seconds)
x_3 = Trunk extension ratio (N/kg)

Suppose that the model equation is

$$y = 30 + 0.90x_1 + 0.08x_2 - 4.50x_3 + e$$

and that $\sigma = 5$.
a. What is the population regression function?
b. What are the values of the population regression coefficients?
c. Interpret the value of β_1.
d. Interpret the value of β_3.

14.4 Consider the regression model introduced in the previous exercise.
a. What is the mean rating of acceptable weight when extent of left lateral bending is 25 cm, dynamic hand grip endurance is 200 seconds, and trunk extension ratio is 10 N/kg?
b. If repeated observations on rating are made on different individuals, all of whom have the values of x_1, x_2, and x_3 specified in Part (a), in the long run approximately what percentage of ratings will be between 13.5 kg and 33.5 kg?

14.5 The authors of the paper **"Predicting Yolk Height, Yolk Width, Albumen Length, Eggshell Weight, Egg Shape Index, Eggshell Thickness, Egg Surface Area of Japanese Quails Using Various Egg Traits as Regressors"** (***International Journal of Poultry Science*** **[2008]: 85–88)** used a multiple regression model with two independent variables where

y = Quail egg weight (g)
x_1 = Egg width (mm)
x_2 = Egg length (mm)

The regression function suggested in the paper is $y = -21.658 + 0.828x_1 + 0.373x_2$.
a. What is the mean egg weight for quail eggs that have a width of 20 mm and a length of 50 mm?
b. Interpret the values of β_1 and β_2.

14.6 Example 14.4 considered using a multiple regression model to estimate the body mass of Chinstrap penguins (**"Palmer Archipelago Penguins Data in the palmerpenguins R Package - An Alternative to Anderson's Irises,"** ***The R Journal,*** **2022)**. Three of the variables studied were body mass, flipper length, and sex. Based on the data from this paper, a reasonable population model that includes the predictors flipper length and sex and that also includes a sex-flipper length interaction variable is

$$y = 1225 - 5340x_1 + 12x_2 + 28.3x_3 + \varepsilon$$

Copyright 2025 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s). Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

where the variables are defined as

y = Body mass (in grams)
x_1 = Sex (1 = male, 0 = female)
x_2 = Flipper length (in mm)
$x_3 = x_1x_2$ (an interaction term)

a. Using the given model, estimate the mean body mass of female chinstrap penguins with a flipper length of 195 mm.
b. How much different is the estimate of body mass for chinstrap penguins with a flipper length of 195 mm who are male compared to the prediction for females with this same flipper length?
c. Interpret the coefficient of the flipper length variable.
d. Explain what it means to include an interaction variable in this context. (Hint: See Example 14.4.)

14.7 The authors of the paper **"Power-Load Prediction Based on Multiple Linear Regression Model" (*Boletin Tecnico* [2017]: 390–397)** were interested in predicting the load on the electric power system in China using data on y = Power consumption (in hundreds of millions of kwh), x_1 = Population (in millions), and x_2 = Gross domestic product (in billions of dollars), for 21 years. The model equation proposed in the paper is

$$y = -113{,}527 + 0.974x_1 + 0.057x_2 + e$$

a. According to this model, what is the mean power consumption for a year if the population was 130,000 million and the gross domestic product was 400,000 billion dollars?
b. Interpret the value of β_1 in this model.

14.8 Groundwater is the main source of water in many countries. In a study of the quality of ground water in Yemen, the authors of the paper **"Multiple Linear Regression Model for Chloride Estimation of Groundwater in Ash-Shihr Town and Its Outskirts Hadharamout-Yemen" (*International Journal of Environmental Sciences* [2016]: 881–888)** used a multiple regression model with two independent variables, where

y = Chloride concentration (mg/L)
x_1 = Total alkalinity
x_2 = Electrical conductivity

The model equation suggested in the paper is

$$y = -183.560 + 1.589x_1 + 0.069x_2 + e$$

a. Based on this model, what is the mean value of y when $x_1 = 300$ and $x_2 = 2500$?
b. Based on this model, what mean chloride concentration is associated with a total alkalinity of 230 and an electrical conductivity of 2700?

14.9 The article **"Effects of Age on the Variability and Stability of Gait: A Cross-Sectional Treadmill Study in Healthy Individuals Between 20 and 69 Years of Age" (*Gait and Posture* [2014]: 170–174)** proposed a quadratic regression model to describe the relationship between a measure of gait instability, y, and age. The model suggested in the paper is

$$y = 0.92 - 0.005x + 0.00007x^2 + e$$

a. Graph the function $y = 0.92 - 0.005 + 0.00007x^2$ over x values between 20 and 70. (Hint: Substitute x = 20, 30, 40, 50, 60, and 70 into the function to find points on the graph and connect them with a smooth curve, or use appropriate technology.)
b. Would the mean gait instability be higher for age 40 or age 60?
c. What is the change in mean gait instability when age increases from 20 to 30? From 50 to 60?

14.10 The relationship between yield of maize (a type of corn), date of planting, and planting density was investigated in the article **"Development of a Model for Use in Maize Replant Decisions" (*Agronomy Journal* [1980]: 459–464)**. Let

y = Maize yield (percent)
x_1 = Planting date (days after April 20)
x_2 = Planting density (10,000 plants/ha)

The regression model with both quadratic terms ($y = \alpha + \beta_1x_1 + \beta_2x_2 + \beta_3x_3 + \beta_4x_4 + e$ where $x_3 = x_1^2$ and $x_4 = x_2^2$) provides a good description of the relationship between y and the independent variables.
a. If $\alpha = 21.09$, $\beta_1 = 0.653$, $\beta_2 = 0.0022$, $\beta_3 = 2.0206$, and $\beta_4 = 0.4$, what is the population regression function?
b. Use the regression function from Part (a) to determine the mean yield for a plot planted on May 6 with a density of 41,180 plants/ha.
c. Would the mean yield be higher for a planting date of May 6 or May 22 (for the same density)?
d. Is it appropriate to interpret $\beta_1 = 0.653$ as the average change in yield when planting date increases by one day and the values of the other three predictors are held fixed? Why or why not?

14.11 Suppose that the variables y, x_1, and x_2 are related by the regression model

$$y = 1.8 + 0.1x_1 + 0.8x_2 + e$$

a. Construct a graph (similar to that of Figure 14.5) showing the relationship between mean y and x_2 for fixed values 10, 20, and 30 of x_1.

Copyright 2025 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s). Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

b. Construct a graph depicting the relationship between mean y and x_1 for fixed values 50, 55, and 60 of x_2.

c. What aspect of the graphs from Parts (a) and (b) can be attributed to the lack of an interaction between x_1 and x_2?

d. Suppose the interaction term $0.03x_3$ where $x_3 = x_1x_2$ is added to the regression model equation. Using this new model, construct the graphs described in Parts (a) and (b). How do they differ from those of Parts (a) and (b)?

14.12 A manufacturer of wood stoves collected data on y = Particulate matter concentration and x_1 = Flue temperature for three different air intake settings (low, medium, and high).

a. Write a model equation that includes indicator variables to incorporate intake setting, and interpret each of the β coefficients.

b. What additional predictors would be needed to incorporate interaction between temperature and intake setting?

14.13 Consider a regression analysis with three independent variables x_1, x_2, and x_3. Give the equation for the following regression models:

a. The model that includes as predictors all independent variables but no quadratic or interaction terms.

b. The model that includes as predictors all independent variables and all quadratic terms.

c. All models that include as predictors all independent variables, no quadratic terms, and exactly one interaction term.

d. The model that includes as predictors all independent variables, all quadratic terms, and all interaction terms (the full quadratic model).

14.14 Animals are often transported long distances, and this can be stressful for them. The authors of the paper **"Prediction of Stress Responses in Goats: Comparison of Artificial Neural Network and Multiple Regression Models" (*Canadian Journal of Animal Science* [2019]: 102–110)** investigated the effect of transportation time on several physiological measures of stress. They developed a quadratic regression model of the relationship between x = Transportation time and y = Blood glucose level (mg/dL) using data from an experiment in which 80 goats were assigned to one of five travel times (0, 1, 2, 3, and 4 hours, with 16 goats in each travel time group). The following regression model was suggested:

$$y = 156.7 + 34.88x - 18.23x^2$$

a. According to this model, what is the mean blood glucose level for a goat that is transported for 3 hours?

b. For this particular model, does it make sense to interpret the value of β_1 as the average change in blood glucose level associated with a 1-hour increase in travel time?

c. Explain why you would not want to use this model to estimate the mean blood glucose level of goats that are transported for 8 hours.

14.15 Consider the dependent variable y = Fuel efficiency of a car (mpg).

a. Suppose that you want to incorporate type of car, with four categories (subcompact, compact, midsize, and large), into a regression model that also includes x_1 = Age of car and x_2 = Engine size. Define the necessary indicator variables, and write out the complete model equation.

b. Suppose that you want to incorporate interaction between age and type of car. What additional predictors would be needed to accomplish this?

Section 14.2 Fitting a Model and Assessing Its Utility

In Section 14.1, multiple regression models containing several different types of predictors were introduced. Now suppose that a particular set of k predictor variables $x_1, x_2, \ldots, x_k$ has been selected for inclusion in the model

$$y = \alpha + \beta_1x_1 + \beta_2x_2 + \cdots + \beta_kx_k + e$$

The next steps are to estimate the model coefficients $\alpha, \beta_1, \ldots, \beta_k$ and the regression function $\alpha + \beta_1x_1 + \cdots + \beta_kx_k$ (the mean y value for specified values of the predictors), assess the model's usefulness, and if appropriate, use the estimated model to make predictions. This requires sample data. As before, n denotes the number of observations in the sample. With just one independent variable, the sample consisted of n (x, y) pairs. Now, each observation consists of $(k + 1)$ numbers: a value of x_1, a value of $x_2, \ldots,$ a value of x_k, and the associated value of y. The n observations are assumed to have been selected independently of one another.

Copyright 2025 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s). Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

Example 14.6 Graduation Rates at Small Colleges

Understand the context ➤

● One way colleges measure success is by graduation rates. **The Education Trust** publishes 6-year graduation rates along with other college characteristics on its website **(collegeresults.org)**. Consider the following variables:

y = 6-year graduation rate
x_1 = Median SAT score of students accepted to the college
x_2 = Student-related expense per full-time student (in dollars)

$$x_3 = \begin{cases} 1 & \text{if college only admits students who identify as women} \\ & \text{or only admits students who identify as men} \\ 0 & \text{if college admits students of all genders} \end{cases}$$

The following data represent a random sample of 22 colleges selected from the 1037 colleges in the United States with enrollments under 5000 students. The data consist of 22 observations on each of these four variables.

College	y	x_1	x_2	x_3
Cornerstone University	0.391	1,065	9,482	0
Barry University	0.389	950	13,149	0
Wilkes University	0.532	1,090	9,418	0
Colgate University	0.893	1,350	26,969	0
Lourdes College	0.313	930	8,489	0
Concordia University at Austin	0.315	985	8,329	0
Carleton College	0.896	1,390	29,605	0
Letourneau University	0.545	1,170	13,154	0
Ohio Valley College	0.288	950	10,887	0
Chadron State College	0.469	990	6,046	0
Meredith College	0.679	1,035	14,889	1
Tougaloo College	0.495	845	11,694	0
Hawaii Pacific University	0.410	1,000	9,911	0
University of Michigan-Dearborn	0.497	1,065	9,371	0
Whittier College	0.553	1,065	14,051	0
Wheaton College	0.845	1,325	18,420	0
Southampton College of Long Island	0.465	1,035	13,302	0
Keene State College	0.541	1,005	8,098	0
Mount St Mary's College	0.579	918	12,999	1
Wellesley College	0.912	1,370	35,393	1
Fort Lewis College	0.298	970	5,518	0
Bowdoin College	0.891	1,375	35,669	0

One possible model that could be considered to describe the relationship between y and these three independent variables is

$$y = \alpha + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + e$$

This example will be revisited after you see how sample data are used to estimate model coefficients.

As in simple linear regression, the principle of least squares is used to estimate the coefficients $\alpha, \beta_1, \ldots, \beta_k$. For specified estimates $a, b_1, \ldots, b_k$

● **Data set available online**

$$y - (a + b_1 x_1 + b_2 x_2 + \cdots + b_k x_k)$$

is the deviation between the observed y value for a particular observation and the predicted value using the estimated regression function $a + b_1x_1 + \cdots + b_kx_k$. For example, the first observation in the data set of Example 14.6 is

$$(y, x_1, x_2, x_3) = (0.391, 1065, 9482, 0)$$

The resulting deviation between observed and predicted y values is

$$0.391 - [a + b_1(1065) + b_2(9482) + b_3(0)]$$

Deviations corresponding to other observations are found in a similar manner. Using the principle of least squares, the estimates of α, β_1, β_2, and β_3 are the values of a, b_1, b_2, and b_3 that minimize the sum of these squared deviations.

According to the principle of least squares, the fit of a particular estimated regression function $a + b_1x_1 + \cdots + b_kx_k$ to the observed data is measured by the sum of the squared deviations between the observed y values and the y values predicted by the estimated regression function:

$$\Sigma[y - (a + b_1x_1 + \cdots + b_kx_k)]^2$$

The **least-squares estimates** of $\alpha, \beta_1, \ldots, \beta_k$ are those values of $a, b_1, \ldots, b_k$ that make this sum of squared deviations as small as possible.

The least-squares estimates for a given data set are obtained by solving a system of $(k + 1)$ equations in the $(k + 1)$ unknowns $a, b_1, \ldots, b_k$ (called the *normal equations*). In the case $k = 1$ (simple linear regression), there are only two equations, and their general solution—the expressions for b and a—was given in Chapter 5. For $k \geq 2$, it is not as easy to write general expressions for the estimates without using advanced mathematical notation. Fortunately all the commonly used statistical software packages can calculate these estimates using sample data.

Example 14.7 More on Graduation Rates at Small Colleges

Understand the context ➤

Figure 14.6 displays Minitab output for the model $y = \alpha + \beta_1x_1 + \beta_2x_2 + \beta_3x_3 + e$ based on the small college data of Example 14.6. Focus on the column labeled Coef (for coefficient) in the table near the top of the figure. The four numbers in this column are the estimated model coefficients:

$a = -0.3906$ (the estimate of the constant term α)
$b_1 = 0.0007602$ (the estimate of the coefficient β_1)
$b_2 = 0.00000697$ (the estimate of the coefficient β_2)
$b_3 = 0.12495$ (the estimate of the coefficient β_3)

Interpret the results ➤

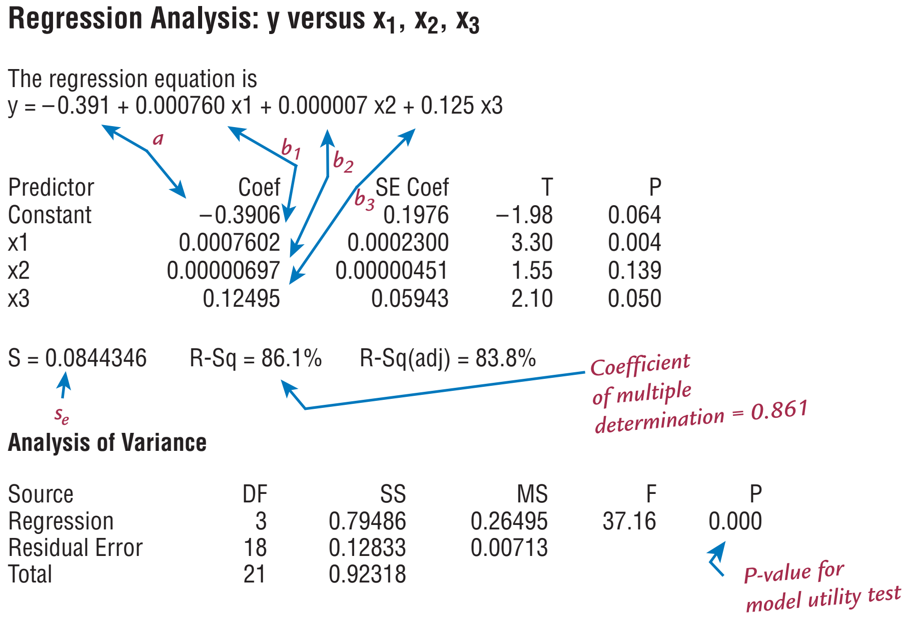
You would estimate that the average change in 6-year graduation rate associated with a \$1 increase in expenditure per full-time student while type of institution (colleges that admit only students who identify as women or only students who identify as men or colleges that admit students of all genders) and median SAT score remains fixed is 0.00000697. A similar interpretation applies to b_1.

The variable x_3 is an indicator variable that takes on a value of 1 for colleges that only admit students who identify as women or only admit students who identify as men. You would interpret the estimated value of $b_3 = 0.125$ as the "correction" that you would make to the predicted 6-year graduation rate of a college that admits students of all genders with the same median SAT and expenditure per full-time student to incorporate the difference associated with having only students who identify as women or only students who identify as men.

The estimated regression function is

$$\left(\begin{array}{c}\text{estimated mean value of } y \\ \text{for specified } x_1, x_2, \text{ and } x_3 \text{ values}\end{array}\right) = -0.3906 + 0.0007602x_1 + 0.00000697x_2 + 0.12495x_3$$

Figure 14.6
Minitab output for the regression analysis of Example 14.7.



```
Regression Analysis: y versus x1, x2, x3

The regression equation is
y = −0.391 + 0.000760 x1 + 0.000007 x2 + 0.125 x3

Predictor           Coef       SE Coef        T        P
Constant         −0.3906        0.1976    −1.98    0.064
x1             0.0007602     0.0002300     3.30    0.004
x2            0.00000697    0.00000451     1.55    0.139
x3               0.12495       0.05943     2.10    0.050

S = 0.0844346    R-Sq = 86.1%    R-Sq(adj) = 83.8%

Analysis of Variance

Source            DF         SS         MS        F        P
Regression         3    0.79486    0.26495    37.16    0.000
Residual Error    18    0.12833    0.00713
Total             21    0.92318
```

a
b_1
b_2
b_3
s_e
Coefficient of multiple determination = 0.861
P-value for model utility test

Substituting $x_1 = 1000$, $x_2 = 11{,}000$, and $x_3 = 0$ gives

$$-0.3906 + 0.0007602(1000) + 0.00000697(11{,}000) + 0.12495(0) = 0.4462$$

which can be interpreted as an estimate for the mean 6-year graduation rate of colleges that admit students of all genders and that have a median SAT of 1000 and an expenditure per full-time student of \$11,000. This is also the predicted graduation rate for a single college with these same characteristics.

Is the Model Useful?

Whether an estimated model is useful can be assessed by examining the extent to which predicted y values based on the estimated regression function are close to the y values actually observed.

The first predicted value $\hat{y}_1$ is obtained by taking the values of the independent variables $x_1, x_2, \ldots, x_k$ for the first sample observation and substituting these values into the estimated regression function.

Doing this successively for the remaining observations results in the **predicted values** $\hat{y}_2, \ldots, \hat{y}_n$.

The **residuals** are then the differences between the observed and predicted y values, $(y_1 - \hat{y}_1), (y_2 - \hat{y}_2), \ldots, (y_n - \hat{y}_n)$.

In multiple regression, the predicted values and residuals are defined in the same way as in simple linear regression, but calculation of the values is more involved because there is more than one predictor. Fortunately, statistical software can be used to find the predicted values (the $\hat{y}$'s) and the residuals (the $(y - \hat{y})$'s).

Consider again the college data from Examples 14.6 and 14.7. Because the first y observation, $y_1 = 0.391$, was for a college with $x_1 = 1065$, $x_2 = 9482$, and $x_3 = 0$, the first predicted value is

$$\hat{y} = -0.3906 + 0.0007602(1065) + 0.00000697(9482) + 0.12495(0) = 0.485$$

The first residual is then

$$(y_1 - \hat{y}_1) = (0.391 - 0.485) = -0.094$$

Copyright 2025 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s). Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

The other predicted values and residuals are calculated in a similar way. The sum of residuals from a least-squares fit should be 0, except for small differences due to rounding.

As in simple linear regression, the sum of squared residuals is the basis for several important summary quantities that tell you about the usefulness of a model.

The **residual (or error) sum of squares, SSResid**, and **total sum of squares, SSTo**, are given by

$$\text{SSResid} = \Sigma(y - \hat{y})^2 \qquad \text{SSTo} = \Sigma(y - \bar{y})^2$$

where $\bar{y}$ is the mean of the y observations in the sample.

The number of degrees of freedom associated with SSResid is $n - (k + 1)$, because $(k + 1)$ df are lost in estimating the $k + 1$ coefficients $\alpha, \beta_1, \ldots, \beta_k$.

An estimate of the random error variance σ^2 is given by

$$s_e^2 = \frac{\text{SSResid}}{n - (k + 1)}$$

and $s_e = \sqrt{s_e^2}$ is an estimate of $\boldsymbol{\sigma}$.

The **coefficient of multiple determination**, R^2, interpreted as the proportion of the variability in observed y values that is explained by the fitted model, is

$$R^2 = 1 - \frac{\text{SSResid}}{\text{SSTo}}$$

Example 14.8 Small Colleges Revisited

Figure 14.6 gives Minitab output for the college data for a three-predictor model. The residual sum of squares is found in the Residual Error row and SS column of the table headed Analysis of Variance: SSResid $= 0.12833$. The associated number of degrees of freedom is $n - (k + 1) = 22 - (3 + 1) = 18$, which appears in the DF column just to the left of SSResid.

The sample mean y value is $\bar{y} = 0.5544$, and $\text{SSTo} = \Sigma(y - 0.5544)^2 = 0.92318$ appears in the Total row and SS column of the Analysis of Variance table just under the value of SSResid. The values of s_e, s_e^2, and R^2 are then

$$s_e^2 = \frac{\text{SSResid}}{n - (k + 1)} = \frac{0.12833}{18} = 0.007$$

(also found in the MS column of the Minitab output)

$$s_e = \sqrt{s_e^2} = \sqrt{0.007} = 0.084$$

(which appears in the Minitab output just above the Analysis of Variance table)

$$R^2 = 1 - \frac{\text{SSResid}}{\text{SSTo}} = 1 - \frac{0.12833}{0.92318} = 1 - 0.139 = 0.861$$

This means that the percentage of variation explained is $100R^2 = 86.1\%$, which appears on the Minitab output as R-Sq = 86.1%.

Because the value of R^2 is large and the value of s_e is not too large, the values of R^2 and s_e suggest that the chosen model may be useful for predicting y using these three predictors.

In general, a useful model is one that results in both a large R^2 value and a small s_e value. However, there is a catch. These two conditions can be achieved by fitting a model that contains a large number of predictors. Such a model might be successful in explaining the variability in y for the data in the sample, but it almost always specifies a relationship that cannot be generalized to the population and that may be unrealistic and difficult to interpret. What you really want is a model that has relatively few predictors whose roles are easily interpreted and that also explains much of the variability in y.

All statistical software packages include R^2 and s_e in their output, and most also give SSResid. In addition, some packages compute a quantity called the **adjusted R^2**:

$$(\text{adjusted } R^2) = 1 - \left[\frac{n-1}{n-(k+1)}\right]\left(\frac{\text{SSResid}}{\text{SSTo}}\right)$$

Because the quantity in square brackets in the formula is greater than 1, the number subtracted from 1 is larger than SSResid/SSTo, so the adjusted R^2 is smaller than R^2. The value of R^2 must be between 0 and 1, but the adjusted R^2 can, on rare occasions, be negative.

If a large R^2 has been achieved by using just a few independent variables, the adjusted R^2 and R^2 values will not differ greatly. However, the adjustment can be substantial when a large number of predictors (relative to the number of observations) have been used or when R^2 itself is small to moderate (which could happen even when there is no relationship between y and the independent variables). In Example 14.7, the adjusted $R^2 = 0.838$, which is not much less than $R^2 = 0.861$ because the model included only two predictor variables and the sample size was 22.

F Distributions

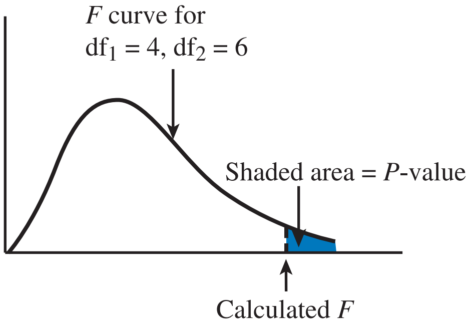
The model utility test in simple linear regression was based on a test statistic that has a t distribution when H_0: $\beta = 0$ is true. The model utility test for multiple regression is based on a test statistic that has a probability distribution called an *F distribution*.

An F distribution arises in connection with a ratio in which the numerator is based on one sum of squares and the denominator is based on a second sum of squares. Each sum of squares has a specified number of degrees of freedom associated with it, so a particular F distribution is determined by specifying values of df_1 = numerator degrees of freedom and df_2 = denominator degrees of freedom. There is a different F distribution for each different df_1 and df_2 combination.

For example, there is an F distribution with 4 numerator degrees of freedom and 12 denominator degrees of freedom, another F distribution with 3 numerator degrees of freedom and 20 denominator degrees of freedom, and so on. A typical F curve for specified numerator and denominator degrees of freedom appears in Figure 14.7.

All F tests introduced in this textbook are upper-tailed. The P-value for an upper-tailed F test is the area under the associated F curve to the right of a calculated F test statistic. Figure 14.7 illustrates this for a test with $\text{df}_1 = 4$ and $\text{df}_2 = 6$.

Figure 14.7
A *P*-value for an upper-tailed *F* test.



If you don't have access to technology, you will need to rely on tables to find P-values for F tests. Unfortunately, tables of these upper-tail areas are more complicated than are tables for t distributions, because two df are involved. The F table (Appendix Table 6) gives only four numbers: the values that capture tail areas 0.10, 0.05, 0.01, and 0.001. Different columns in

Copyright 2025 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s). Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

the table correspond to different values of df_1, and each different group of rows is for a different value of df_2. Figure 14.8 shows how this table is used to obtain *P*-value information.

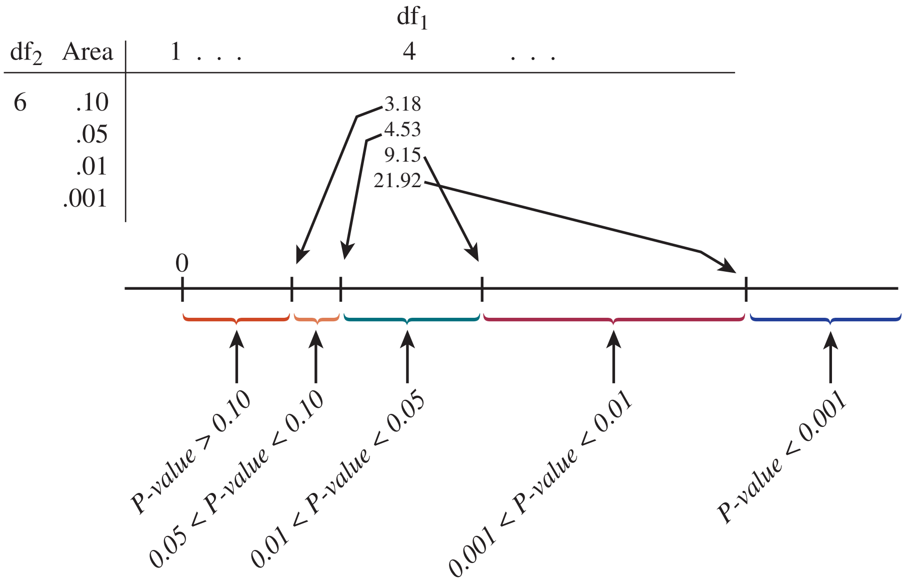


Figure 14.8
Obtaining *P*-value information from the *F* table.

For example, for a test with $\text{df}_1 = 4$ and $\text{df}_2 = 6$,

$$\begin{aligned}&\text{calculated } F = 5.70 \longrightarrow 0.01 < P\text{-value} < 0.05\\&\text{calculated } F = 2.16 \longrightarrow P\text{-value} > 0.10\\&\text{calculated } F = 25.03 \longrightarrow P\text{-value} < 0.001\end{aligned}$$

Only if calculated F equals a tabulated value do you obtain an exact *P*-value from the table (for example, if calculated $F = 4.53$, then $P\text{-value} = 0.05$). If $0.01 < P\text{-value} < 0.05$, you would reject the null hypothesis at a significance level of 0.05 but not at a level of 0.01. When $P\text{-value} < 0.001$, H_0 would be rejected at any reasonable significance level. Statistical software packages such as Minitab, and some graphing calculators can also be used to find *P*-values for F distributions.

The *F* Test for Model Utility

In the simple linear regression model $y = \alpha + \beta x$, if $\beta = 0$, there is no useful linear relationship between y and the single independent variable x. Similarly, if all k coefficients $\beta_1, \beta_2, \ldots, \beta_k$ are 0 in the general k-predictor multiple regression model, there is no useful linear relationship between y and *any* of the independent variables $x_1, x_2, \ldots, x_k$ included in the model. Before using an estimated multiple regression model to make further inferences (for example, predictions or estimates of mean values), it is a good idea to confirm that the model is useful using a formal test procedure.

Recall that SSTo is a measure of total variability in the observed y values and that SSResid measures the amount of total variability that has not been explained by the fitted model. The difference between total and error sums of squares is itself a sum of squares, called the **regression sum of squares**, which is denoted by SSRegr:

$$\text{SSRegr} = \text{SSTo} - \text{SSResid}$$

SSRegr is interpreted as the amount of total variability that *has* been explained by the model. Intuitively, the model should be judged useful if SSRegr is large relative to SSResid and the model uses a small number of predictors relative to the sample size. The number of degrees of freedom associated with SSRegr is k, the number of model predictors, and the number of degrees of freedom for SSResid is $n - (k + 1)$.

The model utility F test is based on the following result:

When all k β_i's are 0 in the model $y = \alpha + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k + e$ and when the distribution of e is normal with mean 0 and variance σ^2 for any particular values of $x_1, x_2, \ldots, x_k$, the statistic

$$F = \frac{\text{SSRegr}/k}{\text{SSResid}/(n - (k + 1))}$$

has an F distribution with numerator $\text{df}_1 = k$ and denominator $\text{df}_2 = n - (k + 1)$.

The value of F tends to be larger when at least one β_i is not 0 than when all the β_i's are 0, because more variability is typically explained by the model. An F statistic value far out in the upper tail of the associated F distribution can be interpreted as evidence that least one β_i is not equal to 0. This is why the model utility F test is upper-tailed.

Model Utility *F* Test for Multiple Regression

Null Hypothesis: H_0: $\beta_1 = \beta_2 = \cdots = \beta_k = 0$
(There is no useful linear relationship between y and *any* of the predictors.)

Alternative Hypothesis: H_a: At least one among $\beta_1, \ldots, \beta_k$ is not zero.
(There is a useful linear relationship between y and *at least one* of the predictors.)

Test Statistic: $$F = \frac{\text{SSRegr}/k}{\text{SSResid}/(n - (k + 1))}$$

where SSRegr = SSTo − SSResid.

An equivalent formula is

$$F = \frac{R^2/k}{(1 - R^2)/(n - (k + 1))}$$

The test is upper-tailed, and technology can be used to find the P-value or the information in Appendix Table 6 can be used to obtain a bound or bounds on the P-value using numerator $\text{df}_1 = k$ and denominator $\text{df}_2 = n - (k + 1)$.

Assumptions: For any particular combination of predictor variable values, the distribution of e, the random deviation, is *normal* with mean 0 and *constant variance,* σ^2.

For the model utility test, the null hypothesis is the claim that the model is not useful. Unless H_0 can be rejected, the model has not demonstrated that it is useful.

Example 14.9 Small Colleges One Last Time

● The model fit to the college data introduced in Example 14.6 included $k = 3$ predictors. The Minitab output in Figure 14.6 contains the relevant information for carrying out the model utility test.

Understand the context ➤

1. The model is $y = \alpha + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + e$ where y = 6-year graduation rate, x_1 = Median SAT score, x_2 = Expenditure per full-time student, and x_3 is an indicator variable that is equal to 1 for a college that admits only students who identify as women or only students who identify as men and is equal to 0 if the college admits students of all genders.
2. H_0: $\beta_1 = \beta_2 = \beta_3 = 0$ (the model is not useful)

● **Data set available online**

Copyright 2025 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s). Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

3. H_a: At least one of the three β_i's is not zero.

4. Significance level: $\alpha = 0.05$

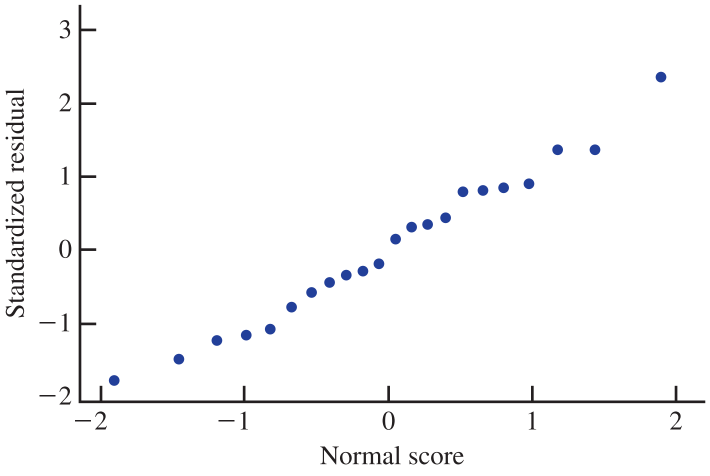
Formulate a plan ➤

5. Test statistic: $F = \dfrac{\text{SSRegr}/k}{\text{SSResid}/[n - (k + 1)]}$

6. Assumptions: The accompanying table gives the residuals and standardized residuals (from Minitab) for the model under consideration.

Observation	***y***	***x*1**	***x*2**	***x*3**	**Residual**	**Standardized Residual**
1	0.391	1,065	9,482	0	−0.094	−1.166
2	0.389	950	13,149	0	−0.034	−0.442
3	0.532	1,090	9,418	0	0.028	0.358
4	0.893	1,350	26,969	0	0.069	0.908
5	0.313	930	8,489	0	−0.062	−0.779
6	0.315	985	8,329	0	−0.101	−1.244
7	0.896	1,390	29,605	0	0.024	0.319
8	0.545	1,170	13,154	0	−0.045	−0.575
9	0.288	950	10,887	0	−0.119	−1.497
10	0.469	990	6,046	0	0.065	0.812
11	0.679	1,035	14,889	1	0.054	0.806
12	0.495	845	11,694	0	0.162	2.388
13	0.410	1,000	9,911	0	−0.029	−0.350
14	0.497	1,065	9,371	0	0.013	0.158
15	0.553	1,065	14,051	0	0.036	0.441
16	0.845	1,325	18,420	0	0.100	1.381
17	0.465	1,035	13,302	0	−0.024	−0.292
18	0.541	1,005	8,098	0	0.111	1.371
19	0.579	918	12,999	1	0.056	0.857
20	0.912	1,370	35,393	1	−0.110	−1.793
21	0.298	970	5,518	0	−0.087	−1.091
22	0.891	1,375	35,669	0	−0.012	−0.189

A normal probability plot of the standardized residuals is shown below. The plot is reasonably straight, indicating that the assumption of normality of the random deviation distribution is plausible.



Copyright 2025 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s). Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

Do the work ➤ **7.** Directly from the table in Figure 14.6, the SS column gives SSRegr = 0.79486 and SSResid = 0.12833. Then

$$F = \frac{0.79486/3}{0.12833/18} = \frac{0.26495}{0.00713} = 37.16 \quad \left(\begin{array}{l}\text{also found in the column labeled F}\\ \text{in Figure 14.6}\end{array}\right)$$

8. Appendix Table 6 shows that for a test with $df_1 = k = 3$ and $df_2 = n - (k + 1) = 22 - (3 + 1) = 18$, the value 8.49 captures upper-tail F curve area 0.001. Since calculated $F = 37.16 > 8.49$, it follows that P-value < 0.001. The P-value for the model utility test (P-value $= 0.000$) can also be found in the Minitab output to the right of the value of the F test statistic in the column labeled P.

Interpret the results ➤ **9.** Because the P-value is less than the significance level of 0.05, H_0 should be rejected. The conclusion would be the same using $\alpha = 0.01$ or $\alpha = 0.001$. The usefulness of the multiple regression model is confirmed.

Example 14.10 Compression Strength of Wood

Understand the context ➤ Wood is sometimes processed using a heat treatment to improve durability and resistance to fungi. A multiple regression analysis presented in the article **"An Artificial Neural Network Model for Predicting Compression Strength of Heat Treated Woods and Comparison with a Multiple Linear Regression Model" (*Construction and Building Materials* [2014]: 102–108)** considered a model in which the dependent variable was

y = Compression strength of heat-treated wood (N/mm^2)

and the predictors were

x_1 = Heat treatment temperature (°C)
x_2 = Heat treatment time (hours)

Summary quantities included $n = 48$ and $R^2 = 0.83$. Does this model specify a useful relationship between y and the two predictors? To answer this question, you can carry out a model utility test:

1. The fitted model was $y = \alpha + \beta_1 x_1 + \beta_2 x_2 + e$
2. H_0: $\beta_1 = \beta_2 = 0$
3. H_a: at least one of the β_i's is not zero
4. Significance level: $\alpha = 0.01$

Formulate a plan ➤ **5.** Test statistic: $F = \dfrac{R^2/k}{(1 - R^2)/(n - (k + 1))}$

6. Assumptions: The raw data were not given in this article, so it isn't possible to compute standardized residuals or construct a normal probability plot. For this test to be valid, you must be willing to assume that the random deviation distribution is normal.

Do the work ➤ **7.** $F = \dfrac{0.83/2}{(0.17)/(48 - (2 + 1))} = \dfrac{0.415}{0.004} = 103.75$

8. The test has $df_1 = k = 2$ and $df_2 = n - (k + 1) = 45$. This latter df is not included in the F table. However, the 0.001 cutoff value for $df_2 = 40$ is 8.25, and for $df_2 = 60$ it is 7.77. The value of the test statistic ($F = 103.75$) is much greater than these values, implying that the P-value < 0.001.

Interpret the results ➤ **9.** Since P-value < 0.001 which is less than the significance level of 0.01, H_0 is rejected at significance level 0.01. There appears to be a useful linear relationship between y and at least one of the two predictors.

Copyright 2025 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s). Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

In Section 14.3 (online), you will see how a model that has been judged to be useful can be used to draw further conclusions. However, it is important to realize that in many applications, more than one model's utility could be confirmed by the F test. Also, just because the model utility test indicates that the multiple regression model is useful does not necessarily mean that all the predictors included in the model contribute to the usefulness of the model. This is illustrated in Example 14.11, and strategies for selecting a model are considered later in Section 14.4 (online).

Example 14.11 The Cost of Black-and-White Printers

Understand the context ➤

What factors contribute to the cost of a black-and-white computer printer? The article **"Top Picks for a Great Home Office" (*Consumer Reports* [May/June 2023]: 20–27)** included data on the following variables for 43 black-and-white printers:

y = Price

x_1 = Ink/toner cost per year for typical use

x_2 = Print speed (pages per minute)

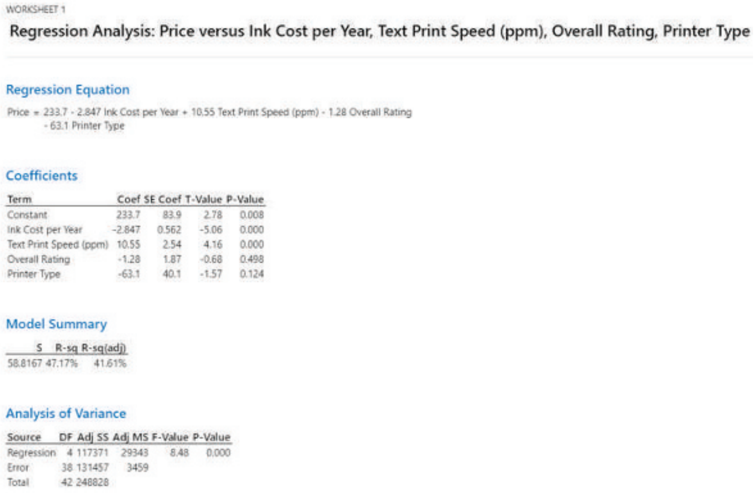
x_3 = *Consumer Reports* overall rating (on a scale of 0 to 100)

x_4 = Printer type (0 for inkjet, 1 for laser)

The data from the article were used to fit the multiple regression model

$$y = \alpha + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + \varepsilon$$

Minitab output is shown in Figure 14.9.



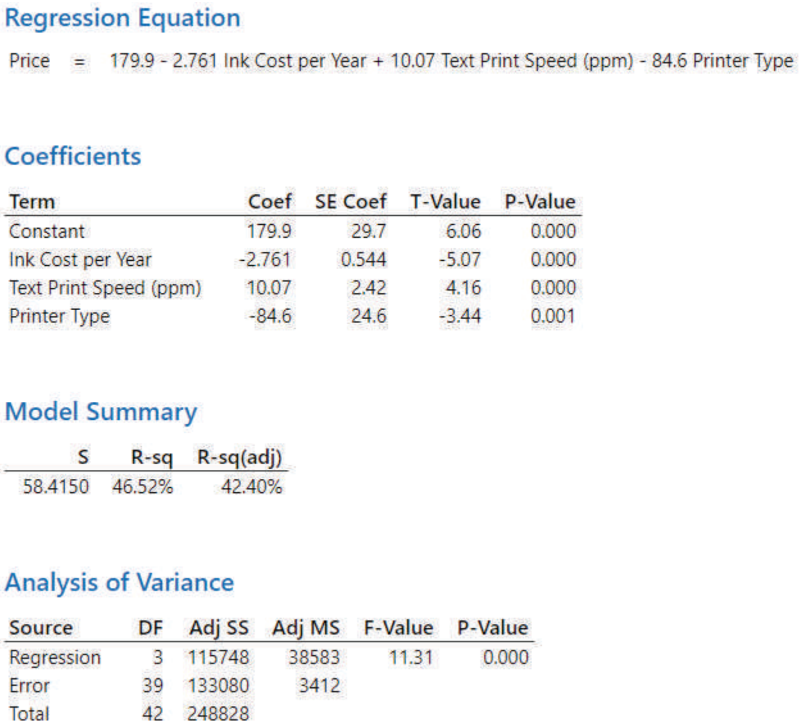
```
WORKSHEET 1
Regression Analysis: Price versus Ink Cost per Year, Text Print Speed (ppm), Overall Rating, Printer Type

Regression Equation
Price = 233.7 - 2.847 Ink Cost per Year + 10.55 Text Print Speed (ppm) - 1.28 Overall Rating
          - 63.1 Printer Type

Coefficients
Term                    Coef  SE Coef  T-Value  P-Value
Constant               233.7     83.9     2.78    0.008
Ink Cost per Year     -2.847    0.562    -5.06    0.000
Text Print Speed (ppm) 10.55     2.54     4.16    0.000
Overall Rating         -1.28     1.87    -0.68    0.498
Printer Type           -63.1     40.1    -1.57    0.124

Model Summary
      S    R-sq  R-sq(adj)
58.8167  47.17%     41.61%

Analysis of Variance
Source      DF  Adj SS  Adj MS  F-Value  P-Value
Regression   4  117371   29343     8.48    0.000
Error       38  131457    3459
Total       42  248828
```

Figure 14.9
Minitab output for the printer data of Example 14.11.

Interpret the results ➤

From the Minitab output, $F = 8.48$, with an associated P-value of 0.000, indicating that the null hypothesis in the model utility test, H_0: $\beta_1 = \beta_2 = \beta_3 = 0$ should be rejected. You would conclude that there is a useful linear relationship between y and at least one of x_1, x_2, x_3, and x_4. However, consider the Minitab output shown in Figure 14.10, which resulted from fitting a model that does not include the variable x_3, which was the *Consumer Report* overall rating. Notice that the model utility F test would also confirm the usefulness of this model and that the adjusted R^2 value is actually even a bit greater for the model with only three predictors (42.40% compared to 41.61%). This suggests that the simpler model with just three predictors may be preferred over the model with four predictors.

Copyright 2025 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s). Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.



```
Regression Equation

Price  =  179.9 - 2.761 Ink Cost per Year + 10.07 Text Print Speed (ppm) - 84.6 Printer Type


Coefficients

Term                       Coef  SE Coef  T-Value  P-Value
Constant                  179.9     29.7     6.06    0.000
Ink Cost per Year        -2.761    0.544    -5.07    0.000
Text Print Speed (ppm)    10.07     2.42     4.16    0.000
Printer Type              -84.6     24.6    -3.44    0.001


Model Summary

      S    R-sq  R-sq(adj)
58.4150  46.52%     42.40%


Analysis of Variance

Source      DF  Adj SS  Adj MS  F-Value  P-Value
Regression   3  115748   38583    11.31    0.000
Error       39  133080    3412
Total       42  248828
```

Figure 14.10
Minitab output for the printer data of Example 14.11 when the variable x_3 is not included as a predictor.

Exercises 14.16–14.28

● **Data set available online**

14.16 Exercise 14.2 described a study investigating factors that influence student satisfaction and perceived learning in online courses (**"Predicting Student Satisfaction and Perceived Learning Within Online Learning Environments," *Distance Education* [2019]: 133–148**). Data were collected from a sample of 167 college students who took an online course. They used a multiple regression model to describe the relationship between the following variables

y = A measure of the student's perception of how much they learned, with values ranging from 1 to 5
x_1 = A measure of student-content interaction, with values ranging from 1 to 5
x_2 = A measure of student-instructor interaction, with values ranging from 1 to 5

The authors reported the following for the estimated regression equation $\hat{y} = 1.137 + 0.459x_1 + 0.310x_2$:

$R^2 = 0.400 \quad R^2adj = 0.389 \quad F = 36.268$
$P\text{-value} < 0.001$

a. Interpret the value of b_1.
b. What proportion of the observed variability in the measure of perceived learning can be explained by the model relationship?
c. How does the value of adjusted R^2 compare to the value of R^2?

14.17 State as much information as you can about the P-value for an upper-tailed F test in each of the following situations. (Hint: See the section on F distributions.)

a. $df_1 = 3$, $df_2 = 15$, calculated $F = 4.23$
b. $df_1 = 4$, $df_2 = 18$, calculated $F = 1.95$
c. $df_1 = 5$, $df_2 = 20$, calculated $F = 4.10$
d. $df_1 = 4$, $df_2 = 35$, calculated $F = 4.58$

14.18 Give as much information as you can about the P-value for the model utility F test in each of the following situations:

a. $k = 2$, $n = 21$, calculated $F = 2.47$
b. $k = 8$, $n = 25$, calculated $F = 5.98$
c. $k = 5$, $n = 26$, calculated $F = 3.00$
d. The full quadratic model based on x_1 and x_2 is fit, $n = 20$, and calculated $F = 8.25$.
e. $k = 5$, $n = 100$, calculated $F = 2.33$

14.19 Data from a sample of $n = 150$ quail eggs were used to fit a multiple regression model relating

y = Eggshell surface area (mm^2)
x_1 = Egg weight (g)
x_2 = Egg width (mm)
x_3 = Egg length (mm)

("Predicting Yolk Height, Yolk Width, Albumen Length, Eggshell Weight, Egg Shape Index, Eggshell

Thickness, Egg Surface Area of Japanese Quails Using Various Egg Traits as Regressors," *International Journal of Poultry Science* [2008]: 85–88).

The resulting estimated regression function was

$$\hat{y} = 10.561 + 1.535x_1 - 0.178x_2 - 0.045x_3$$

and $R^2 = 0.996$.

a. Carry out a model utility test to determine if this multiple regression model is useful. (*Hint:* See Example 14.10.)

b. A simple linear regression model was also used to describe the relationship between y and x_1, resulting in the estimated regression function $\hat{y} = 6.254 + 1.387x_1$. The P-value for the associated model utility test was reported to be less than 0.01, and $r^2 = 0.994$. Is the linear model useful? Explain. (Hint: See Example 14.11.)

c. Based on your answers to Parts (a) and (b), which of the two models would you recommend for predicting eggshell surface area? Explain the rationale for your choice.

14.20 ● This exercise requires the use of a statistical software package. The paper **"Habitat Selection by Black Bears in an Intensively Logged Boreal Forrest" (*Canadian Journal of Zoology* [2008]: 1307–1316)** gave the accompanying data on $n = 11$ female black bears.

Age (years)	Weight (kg)	Home Range Size (km^2)
10.5	54	43.1
6.5	40	46.6
28.5	62	57.4
6.5	55	35.6
7.5	56	62.1
6.5	62	33.9
5.5	42	39.6
7.5	40	32.2
11.5	59	57.2
9.5	51	24.4
5.5	50	68.7

a. Fit a multiple regression model to describe the relationship between y = Home range size and the predictors x_1 = Age and x_2 = Weight.

b. Construct a normal probability plot of the 11 standardized residuals. Based on the plot, does it seem reasonable to regard the random deviation distribution as approximately normal? Explain. (Hint: See Example 14.9.)

c. If appropriate, carry out a model utility test with a significance level of 0.05 to determine if at least one of the predictors *age* and *weight* is useful for predicting home range size.

14.21 The ability of ecologists to identify regions of greatest species richness could have an impact on the preservation of genetic diversity, a major objective of the World Conservation Strategy. The article **"Prediction of Rarities from Habitat Variables: Coastal Plain Plants on Nova Scotian Lakeshores" (*Ecology* [1992]: 1852–1859)** used a sample of $n = 37$ lakes to obtain the estimated regression equation

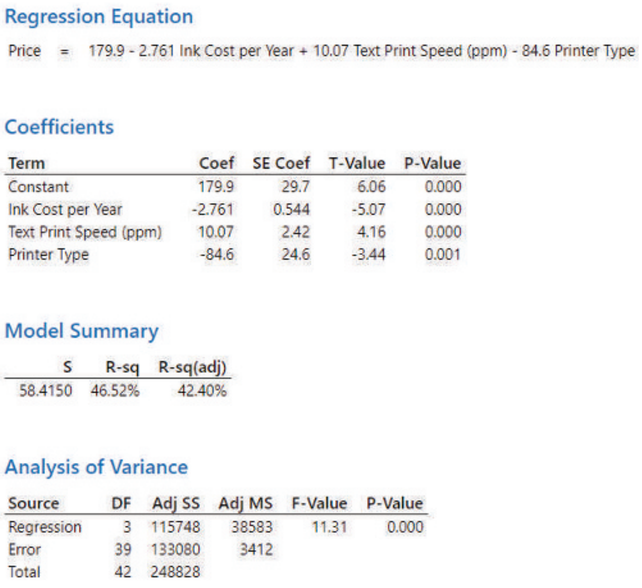
$$\hat{y} = 3.89 + 0.033x_1 + 0.024x_2 + 0.023x_3 + 0.008x_4 - 0.13x_5 - 0.72x_6$$

where y = Species richness, x_1 = Watershed area, x_2 = Shore width, x_3 = Drainage (%), x_4 = Water color (total color units), x_5 = Sand (%), and x_6 = Alkalinity. The coefficient of multiple determination was reported as $R^2 = 0.83$. Use a test with significance level 0.01 to decide whether the chosen model is useful.

14.22 Example 14.11 **("Top Picks for a Great Home Office," *Consumer Reports* [May/June 2023]: 20–27)** included data on the following variables for 43 black-and-white printers:

y = Price
x_1 = Ink/toner cost per year for typical use
x_2 = Print speed (pages per minute)
x_3 = Printer type (0 for inkjet, 1 for laser)

Use the accompanying Minitab output to carry out a test to determine if there is a useful relationship between price and at least one of these predictors.



```
Regression Equation

Price  =  179.9 - 2.761 Ink Cost per Year + 10.07 Text Print Speed (ppm) - 84.6 Printer Type


Coefficients

Term                      Coef   SE Coef   T-Value   P-Value
Constant                 179.9      29.7      6.06     0.000
Ink Cost per Year       -2.761     0.544     -5.07     0.000
Text Print Speed (ppm)   10.07      2.42      4.16     0.000
Printer Type             -84.6      24.6     -3.44     0.001


Model Summary

      S     R-sq   R-sq(adj)
58.4150   46.52%      42.40%


Analysis of Variance

Source      DF   Adj SS   Adj MS   F-Value   P-Value
Regression   3   115748    38583     11.31     0.000
Error       39   133080     3412
Total       42   248828
```

14.23 Refer to the description of variables and the Minitab output in the previous exercise.

a. Interpret the value of b_3, the coefficient of printer type.

b. What proportion of the variability in printer price can be explained by the model that uses ink cost per year, text print speed, and printer type as predictor variables?

Copyright 2025 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s). Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

14.24 ● This exercise requires the use of a statistical software package. *Consumer Reports* **(consumer-reports.org, July 2023)** studied 35 sunscreen lotions and assigned each one an overall score based on its performance in terms of sun protection, protection against ultraviolet rays, and the difference between the advertised sun protection factor (SPF) and the SPF as measured in laboratory testing. The article also reported the price per ounce (in dollars) for each of the sunscreens. The accompanying table gives the data from the article on the following variables:

y = Price per ounce
x_1 = Overall *Consumer Reports* score
x_2 = Label SPF
x_3 = Tested SPF

Price per Ounce	Overall Score	Label SPF	Tested SPF
1.13	100	50	64
5.15	83	50	55
7.60	73	60	33
0.91	68	50	31
0.58	64	50	31
1.57	63	50	32
0.61	61	50	29
0.63	60	50	28
1.25	57	50	20
0.67	56	50	26
8.28	55	30	33
6.33	55	30	16
6.21	55	30	16
5.23	55	60	25
4.00	50	50	19
3.67	50	60	21
1.00	47	30	12
1.38	47	50	16
4.83	46	50	14
1.25	46	30	11
6.21	45	40	11
4.00	43	30	16
2.31	42	30	9
11.99	38	50	12
4.83	36	50	10
3.67	36	30	7
1.27	36	50	11
1.67	35	50	22
1.83	31	50	12
4.00	29	50	14
1.82	28	50	7
2.17	27	50	8
10.50	25	50	9
3.33	18	50	13
3.60	18	30	4

(continued)

a. Fit a multiple regression model to describe the relationship between y = Price per ounce, and the three predictors—overall score, label SPF, and tested SPF.

b. Based on the value of R^2, do you think that the multiple regression model is useful?

c. Are the assumptions for the model utility F test met? (Hint: See Example 14.9.)

14.25 ● This exercise requires the use of a statistical software package. The authors of the article **"Absolute Versus per Unit Body Length Speed of Prey as an Estimator of Vulnerability to Predation" (*Animal Behaviour* [1999]: 347–352)** found that the speed of a prey (twips/s) and the length of a prey (twips × 100) are good predictors of the time (seconds) required to catch the prey. (A twip is a measure of distance used by programmers.) Data were collected in an experiment in which subjects were asked to "catch" an animal of prey moving across their computer screen by clicking on it with the mouse. The investigators varied the length of the prey and the speed with which the prey moved across the screen.

The following data are consistent with summary values and a graph given in the article. Each catch time value represents the average catch time over all subjects. The order of the various speed-length combinations was randomized for each subject.

Prey Length	Prey Speed	Catch Time
7	20	1.10
6	20	1.20
5	20	1.23
4	20	1.40
3	20	1.50
3	40	1.40
4	40	1.36
6	40	1.30
7	40	1.28
7	80	1.40
6	60	1.38
5	80	1.40
7	100	1.43
6	100	1.43
7	120	1.70
5	80	1.50
3	80	1.40
6	100	1.50
3	120	1.90

Copyright 2025 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s). Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

a. Fit a multiple regression model for predicting catch time using prey length and speed as predictors.

b. Predict the catch time for an animal of prey whose length is 6 and whose speed is 50.

c. Is the multiple regression model useful for predicting catch time? Test the relevant hypotheses using $\alpha = 0.05$.

d. The authors of the article suggest that a simple linear regression model with the single predictor

$$x = \frac{\text{length}}{\text{speed}}$$

might be a better model for predicting catch time. Calculate these x values and use them to fit a simple linear regression model.

e. Which of the two models considered (the multiple regression model from Part (a) or the simple linear regression model from Part (d)) would you recommend for predicting catch time? Justify your choice.

14.26 ● This exercise requires the use of a statistical software package. The article **"Vital Dimensions in Volume Perception: Can the Eye Fool the Stomach?"** (*Journal of Marketing Research* [1999]: 313–326) gave the data, shown in the table at the bottom of the page, on dimensions of 27 representative food products.

a. Fit a multiple regression model for predicting the volume (in mL) of a package based on its minimum width, maximum width, and elongation score.

b. Why should you consider adjusted R^2 instead of R^2 when evaluating this model?

c. Carry out a model utility F test.

14.27 The paper referenced in Exercise 14.9 used a quadratic regression model to describe the relationship between y = Gait instability and x = Age. The sample size in this study was $n = 100$ and $R^2 = 0.15$. Carry out a model utility test for the quadratic model. Use $\alpha = 0.05$.

14.28 Suppose that a multiple regression data set consists of $n = 15$ observations. For what values of k, the number of model predictors, would the corresponding model with $R^2 = 0.90$ be judged useful at significance level 0.05? Does such a large R^2 value necessarily imply a useful model? Explain.

Table for exercise 14.26

Product	Material	Height	Maximum Width	Minimum Width	Elongation	Volume
1	glass	7.7	2.50	1.80	1.50	125
2	glass	6.2	2.90	2.70	1.07	135
3	glass	8.5	2.15	2.00	1.98	175
4	glass	10.4	2.90	2.60	1.79	285
5	plastic	8.0	3.20	3.15	1.25	330
6	glass	8.7	2.00	1.80	2.17	90
7	glass	10.2	1.60	1.50	3.19	120
8	plastic	10.5	4.80	3.80	1.09	520
9	plastic	3.4	5.90	5.00	0.29	330
10	plastic	6.9	5.80	4.75	0.59	570
11	tin	10.9	2.90	2.80	1.88	340
12	plastic	9.7	2.45	2.10	1.98	175
13	glass	10.1	2.60	2.20	1.94	240
14	glass	13.0	2.60	2.60	2.50	240
15	glass	13.0	2.70	2.60	2.41	360
16	glass	11.0	3.10	2.90	1.77	310
17	cardboard	8.7	5.10	5.10	0.85	635
18	cardboard	17.1	10.20	10.20	0.84	1,250
19	glass	16.5	3.50	3.50	2.36	650
20	glass	16.5	2.70	1.20	3.06	305
21	glass	9.7	3.00	1.70	1.62	315
22	glass	17.8	2.70	1.75	3.30	305
23	glass	14.0	2.50	1.70	2.80	245
24	glass	13.6	2.40	1.20	2.83	200
25	plastic	27.9	4.40	1.20	3.17	1,205
26	tin	19.5	7.50	7.50	1.30	2,330
27	tin	13.8	4.25	4.25	1.62	730

Chapter Activity

Activity 14.1 Exploring the Relationship Between Number of Predictors and Sample Size

The purpose of this activity is to provide an opportunity to explore what happens when the number of predictors in a multiple regression model is large relative to the sample size.

This activity requires the use of a statistical software package capable of fitting multiple regression models.

Background: The given data on y, x_1, x_2, x_3, and x_4 were generated using a software package capable of producing random observations from any specified normal distribution. Because the data were generated at random, there is no reason to believe that y is related to any of the proposed predictor variables x_1, x_2, x_3, and x_4.

y	x_1	x_2	x_3	x_4
20.5	18.6	22.0	17.1	18.5
20.1	23.9	19.1	21.1	21.3
20.0	20.9	20.7	19.4	20.6
21.7	18.7	18.1	20.9	18.1
20.7	21.1	21.7	23.7	17.0

1. Construct four scatterplots—one of y versus each of x_1, x_2, x_3, and x_4. Do the scatterplots look the way you would expect based on the way the data were generated? Explain.
2. Fit each of the following regression models:
   - **i.** y with x_1
   - **ii.** y with x_1 and x_2
   - **iii.** y with x_1 and x_2 and x_3
   - **iv.** y with x_1 and x_2 and x_3 and x_4
3. Make a table that gives the R^2, the adjusted R^2, and s_e values for each of the models fit in Step 2. Describe what happens to each of these three quantities as additional predictor variables are added to the multiple regression model.
4. Given the manner in which these data were generated, what is the implication of what was observed in Step 3? What does this suggest about the relationship between number of predictors and sample size?

Summary Key Concepts and Formulas

Term or Formula	Comment
Additive multiple regression model, $y = \alpha + \beta_1x_1 + \beta_2x_2 + \cdots + \beta_kx_k + e$	This equation specifies a general probabilistic relationship between y and k independent variables $x_1, x_2, \ldots, x_k$, where $\alpha, \beta_1, \ldots, \beta_k$ are population regression coefficients and $\alpha + \beta_1x_1 + \beta_2x_2 + \cdots + \beta_kx_k$ is the population regression function (the mean value of y for fixed values of $x_1, x_2, \ldots, x_k$).
Estimated regression function, $\hat{y} = a + b_1x_1 + b_2x_2 + \cdots + b_kx_k$	The estimates $a, b_1, \ldots, b_k$ of $\alpha, \beta_1, \ldots, \beta_k$ result from applying the principle of least squares.
Coefficient of multiple determination, $R^2 = 1 - \frac{\text{SSResid}}{\text{SSTo}}$	The proportion of observed y variation that can be explained by the model relationship, where SSResid is defined as it is in simple linear regression but is now based on $n - (k + 1)$ degrees of freedom.
Adjusted R^2	A downward adjustment of R^2 that depends on the number of predictors k and the sample size n.
***F* distribution**	A probability distribution used in the multiple regression model utility test. An F distribution is specified by a numerator df_1 and a denominator df_2.
$F = \frac{\text{SSRegr}/k}{\text{SSResid}/(n - (k + 1))}$ **or** $F = \frac{R^2/k}{(1 - R^2)/(n - (k + 1))}$	The test statistic for testing $H_0\colon \beta_1 = \beta_2 = \cdots = \beta_k = 0$, which states that there is no useful linear relationship between y and any of the model predictors. The F test is upper-tailed and is based on numerator $\text{df}_1 = k$ and denominator $\text{df}_2 = n - (k + 1)$.

Copyright 2025 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s). Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

Chapter 15

Analysis of Variance

isitsharp/iStock/Getty Images

In Chapter 11 methods were introduced for testing H_0: $\mu_1 - \mu_2 = 0$ (that is, $\mu_1 = \mu_2$), where μ_1 and μ_2 are the means of two different populations or the mean responses when two different treatments are applied. Many investigations involve a comparison of more than two population or treatment means.

For example, an investigation was carried out to study the possible consequences of head injuries among soccer players. Ball-to-head impacts are common in soccer because soccer players regularly use their heads to pass the ball and shoot the ball. This is called "heading the ball" and has been associated with head injuries. The authors of the paper **"Effects of Soccer Ball Inflation Pressure and Velocity on Peak Linear and Rotational Accelerations of Ball-To-Head Impacts" (*Sports Engineering* [2020]: 16)** wondered if lower ball inflation pressures would reduce the effect of a ball-to-head impact. They carried out an experiment to compare five different inflation pressures—34, 48, 55, 62, and 76 kPa (kPa, short for kilopascal, is a unit of pressure similar to pounds per square inch). A ball launcher was used to launch the ball at a fixed velocity at the forehead of a robot that could measure the linear acceleration of the impact.

The accompanying table gives approximate means and standard deviations (approximate values based on a graph that appeared in the paper) for peak linear acceleration for balls launched at a speed of 22.2 meters per second.

Sample	Sample Mean Peak Linear Acceleration	Sample Standard Deviation
Inflation Pressure 34 kPa	36.0	0.73
Inflation Pressure 48 kPa	38.0	0.72
Inflation Pressure 55 kPa	39.0	1.04
Inflation Pressure 62 kPa	40.0	0.80
Inflation Pressure 76 kPa	39.5	0.98

Notice that the five sample means are different. But even if the treatment means were equal, you would not expect the five sample means to be exactly equal. Are the differences in the sample means consistent with what is expected simply due to chance differences from one sample to another when the treatment means are equal, or are the differences large enough that you should conclude that the five treatment means are not all equal. This is the type of problem considered in this chapter.

Learning Objectives

After completing this chapter, you will be able to:

- Determine whether there is evidence that two or more population or treatment means are not all equal by conducting a hypothesis test and interpreting the results in context.
- Identify differences in population or treatment means using a multiple comparisons procedure.

Section 15.1 Single-Factor ANOVA and the *F* Test

When two or more populations or treatments are being compared, the characteristic that distinguishes the populations or treatments from one another is called the **factor** under investigation. For example, an experiment might be carried out to compare three different methods for teaching reading (three different treatments), in which case the factor of interest would be *teaching method*, a qualitative factor. If the growth of fish raised in water having different salinity levels—0%, 10%, 20%, and 30%—is of interest, then the factor *salinity level* is quantitative.

A **single-factor analysis of variance (ANOVA)** problem involves a comparison of k population or treatment means $\mu_1, \mu_2, \ldots, \mu_k$. The objective is to test

$$H_0\colon\ \mu_1 = \mu_2 = \cdots = \mu_k$$

against

$$H_a\colon\ \text{At least two of the } \mu\text{'s are different from each other}$$

What Do You Think?

The alternate hypothesis for a single factor ANOVA F test is

$$H_a\colon\ \text{At least two of the } \mu\text{'s are different}$$

How is this different from a hypothesis that states $\mu_1 \neq \mu_2 \neq \cdots \neq \mu_k$?

When comparing populations, the analysis is based on independently selected random samples, one from each population. When comparing treatment means, the data typically result from an experiment and the analysis assumes random assignment of the experimental units (subjects or objects) to treatments. If, in addition, the experimental units are chosen at random from a population of interest, it is also possible to generalize the results of the analysis to this population. (See Chapter 2 for a more detailed discussion of conclusions that can reasonably be drawn based on data from an experiment.)

Whether the null hypothesis in a single-factor ANOVA should be rejected depends on how much the samples from the different populations or treatments differ from one another. Figure 15.1 displays observations that might result when random samples are selected from each of three populations. Each dotplot displays five observations from the first population (using green dots), four observations from the second population (using red dots), and six observations from the third population (using blue dots). For both displays, the three sample means are located by arrows. Notice that the two sample means labeled mean of sample 1 are equal, as are those for the two samples labeled Population 2 and those labeled Population 3.

After looking at the data in Figure 15.1(a), many people would agree that the claim $\mu_1 = \mu_2 = \mu_3$ appears to be false. Not only are the three sample means different, but the three samples are also clearly separated. In this situation, differences between the three sample means are quite large relative to the variability within each sample.

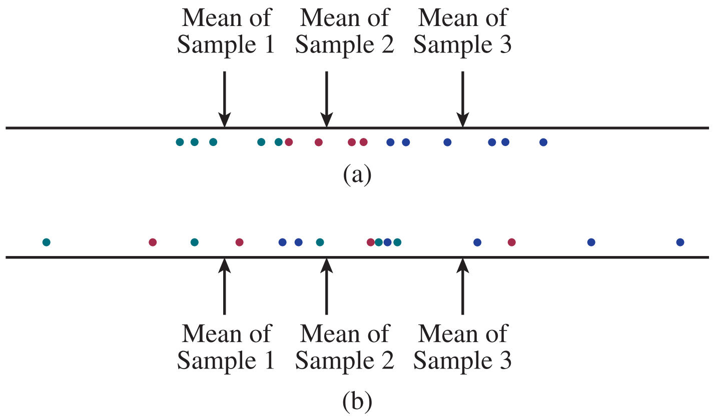


Figure 15.1
Two possible ANOVA data sets when three populations are under investigation:
green dot = observation from Population 1;
red dot = observation from Population 2;
blue dot = observation from Population 3.

The situation pictured in Figure 15.1(b) is much less clear-cut. The sample means are as different as they were in for the samples shown in Figure 15.1(a), but now there is considerable overlap among the three samples. The separation between sample means here might plausibly be attributed to variability in the populations (and therefore the samples) rather than to differences between μ_1, μ_2, and μ_3.

The phrase *analysis of variance* comes from the idea of analyzing variability in the data to see how much can be attributed to differences in the μ's and how much is due to variability in the individual populations. In Figure 15.1(a), the within-sample variability is small relative to the between-sample variability, whereas in Figure 15.1(b), a great deal more of the total variability is due to variability within each sample. If differences between the sample means can be explained by within-sample variability, there is no compelling reason to reject H_0.

Notation and Assumptions

Notation in single-factor ANOVA is a natural extension of the notation used in Chapter 11 for comparing two population or treatment means.

ANOVA Notation

k = number of populations or treatments being compared

Population or treatment	1	2	. . .	k
Population or treatment mean	μ_1	μ_2	. . .	μ_k
Population or treatment variance	σ_1^2	σ_2^2	. . .	σ_k^2
Sample size	n_1	n_2	. . .	n_k
Sample mean	$\bar{x}_1$	$\bar{x}_2$	. . .	$\bar{x}_k$
Sample variance	s_1^2	s_2^2	. . .	s_k^2

$N = n_1 + n_2 + \cdots + n_k$ (the total number of observations in the data set)

T = grand total = sum of all N observations in the data set
$= n_1\bar{x}_1 + n_2\bar{x}_2 + \cdots + n_k\bar{x}_k$

$$\bar{\bar{x}} = \text{grand mean} = \frac{T}{N}$$

A decision between H_0 and H_a is based on examining the $\bar{x}$ values to see whether observed differences are small enough to be attributable simply to sampling variability or whether an alternative explanation for the differences is more plausible.

Example 15.1 Ball-to-Head Impact in Soccer

Understand the context ➤

The chapter introduction described an experiment investigating ball-to-head impacts in soccer **("Effects of Soccer Ball Inflation Pressure and Velocity on Peak Linear and Rotational Accelerations of Ball-To-Head Impacts," *Sports Engineering* [2020]: 16)**. The researchers wondered if lower ball inflation pressures would reduce the effect of a ball-to-head impact. They carried out an experiment to compare five different inflation pressures—34, 48, 55, 62, and 76 kPa (kPa, short for kilopascal, is a unit of pressure similar to pounds per square inch). A ball launcher was used to launch the ball at a fixed velocity at the forehead of a robot that could measure the linear acceleration of the impact.

The accompanying table includes summary values approximated from a graph that appeared in the paper.

Table 15.1 Summary Values for the Ball-to-Head Impact Data of Example 15.1

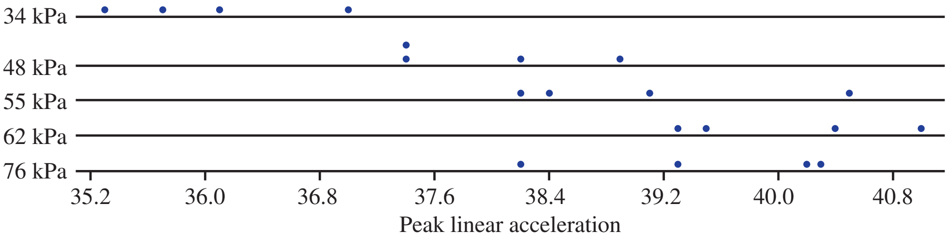
Group	Sample	Sample Size	Sample Mean Peak Linear Acceleration	Sample Standard Deviation
1	Inflation Pressure 34 kPa	4	36.0	0.73
2	Inflation Pressure 48 kPa	4	38.0	0.72
3	Inflation Pressure 55 kPa	4	39.0	1.04
4	Inflation Pressure 62 kPa	4	40.0	0.80
5	Inflation Pressure 76 kPa	4	39.5	0.98

You can use μ_1 to denote the true mean peak linear acceleration for balls inflated to 34 kPa, μ_2 to denote the mean for balls inflated to 48 kPa, and μ_3, μ_4, and μ_5 to represent the other three treatment means. Consider the null hypothesis

$$H_0: \mu_1 = \mu_2 = \mu_3 = \mu_4 = \mu_5$$

Figure 15.2 shows a comparative dotplot for the five samples (based on data consistent with summary values given in Table 15.1).

Figure 15.2
Dotplots for the Data of Example 15.1.



The peak linear acceleration values are lower for the 34 kPa inflation level than for the other inflation levels, but there is quite a bit of overlap in the values for the other four inflation levels. It appears that the null hypothesis of equal means should be rejected, but it is helpful to have a formal test procedure to confirm that this is the correct conclusion.

As with the inferential methods of previous chapters, the validity of the ANOVA test for $H_0: \mu_1 = \mu_2 = \cdots = \mu_k$ requires some assumptions.

Assumptions for ANOVA

1. Each of the k population or treatment response distributions is normal.
2. $\sigma_1 = \sigma_2 = \cdots = \sigma_k$ (The k normal distributions have equal standard deviations.)
3. The observations in the sample from any particular one of the k populations or treatments are independent of one another.
4. When comparing population means, the samples are independent random samples. When comparing treatment means, experimental units are assigned at random to treatments.

Copyright 2025 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s). Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

In practice, the test based on these assumptions works well as long as the assumptions are not too badly violated. If the sample sizes are reasonably large, normal probability plots or boxplots of the data in each sample are helpful for checking the assumption of normality. Often, however, sample sizes are so small that a separate normal probability plot or boxplot for each sample is of little value in checking normality. In this case, a single combined plot can be constructed by first subtracting $\bar{x}_1$ from each observation in the first sample, $\bar{x}_2$ from each value in the second sample, and so on and then constructing a normal probability or boxplot of all N deviations from their respective means. The normal probability plot should be reasonably straight or the boxplot should be approximately symmetric with no outliers. Figure 15.3 shows a boxplot for the data of Example 15.1.

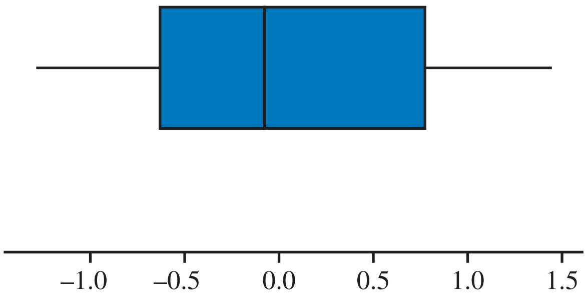


Figure 15.3
A boxplot using the combined data of Example 15.1.

There is a formal procedure for testing the equality of population standard deviations. Unfortunately, it is quite sensitive to even small departures from the normality assumption required for this test, so its use is not recommended. The ANOVA *F* test (to be described later in this section) can safely be used if the largest of the sample standard deviations is at most twice the smallest one. The largest standard deviation in Example 15.1 is $s_3 = 1.04$, which is only about 1.4 times the smallest standard deviation ($s_2 = 0.72$).

The analysis of variance test procedure is based on the following measures of variability in the data.

A measure of differences among the sample means is the **treatment sum of squares**, denoted by **SSTr** and given by

$$\text{SSTr} = n_1(\bar{x}_1 - \bar{\bar{x}})^2 + n_2(\bar{x}_2 - \bar{\bar{x}})^2 + \cdots + n_k(\bar{x}_k - \bar{\bar{x}})^2$$

A measure of variability within the k samples, called **error sum of squares** and denoted by **SSE**, is

$$\text{SSE} = (n_1 - 1)s_1^2 + (n_2 - 1)s_2^2 + \cdots + (n_k - 1)s_k^2$$

Each sum of squares has an associated df:

$$\text{treatment df}_1 = k - 1 \qquad \text{error df}_2 = N - k$$

A **mean square** is a sum of squares divided by its df. In particular,

$$\textbf{mean square for treatments} = \textbf{MSTr} = \frac{\text{SSTr}}{k-1}$$

$$\textbf{mean square for error} = \textbf{MSE} = \frac{\text{SSE}}{N-k}$$

The number of error degrees of freedom comes from adding the number of degrees of freedom associated with each of the sample variances:

$$\begin{aligned}(n_1 - 1) + (n_2 - 1) + \cdots + (n_k - 1) &= n_1 + n_2 + \cdots + n_k - 1 - 1 - \cdots - 1\\ &= N - k\end{aligned}$$

Example 15.2 Ball-to-Head Impact Calculations

For the peak linear acceleration data of Example 15.1, the grand mean $\bar{\bar{x}}$ was calculated to be 38.5. Notice that because the sample sizes are all equal, the grand mean is just the average of the five sample means (this will not usually be the case when the sample sizes are unequal).

With $\bar{x}_1 = 36.0, \bar{x}_2 = 38.0, \bar{x}_3 = 39.0, \bar{x}_4 = 40.0, \bar{x}_5 = 39.5$ and $n_1 = n_2 = n_3 = n_4 = n_5 = 4$

$$\begin{aligned}\text{SSTR} &= n_1(\bar{x}_1 - \bar{\bar{x}})^2 + n_2(\bar{x}_2 - \bar{\bar{x}})^2 + \cdots + n_k(\bar{x}_k - \bar{\bar{x}})^2\\ &= 4(36 - 38.5)^2 + 4(38 - 38.5)^2 + 4(39 - 38.5)^2 + 4(40 - 38.5)^2\\ &\quad + 4(39.5 - 38.5)^2\\ &= 25.0 + 1.0 + 1.0 + 9.0 + 4.0\\ &= 40.0\end{aligned}$$

Because $s_1 = 0.73$, $s_2 = 0.72$, $s_3 = 1.04$, $s_4 = 0.80$, and $s_5 = 0.98$

$$\begin{aligned}\text{SSE} &= (n_1 - 1)s_1^2 + (n_2 - 1)s_2^2 + \cdots + (n_k - 1)s_k^2\\ &= (4 - 1)(0.73)^2 + (4 - 1)(0.72)^2 + (4 - 1)(1.04)^2 + (4 - 1)(0.80)^2\\ &\quad + (4 - 1)(0.98)^2\\ &= 11.200\end{aligned}$$

The numbers of degrees of freedom are

treatment $\text{df}_1 = k - 1 = 4$ error $\text{df}_2 = N - k = 4 + 4 + 4 + 4 + 4 - 4 = 16$

from which

$$\text{MSTr} = \frac{\text{SSTr}}{k - 1} = \frac{40.0}{4} = 10.0$$

$$\text{MSE} = \frac{\text{SSE}}{N - k} = \frac{11.200}{16} = 0.700$$

Both MSTr and MSE are statistics whose values can be calculated once sample data are available. Each of these statistics varies in value from data set to data set. Both of the statistics MSTr and MSE have sampling distributions, and these sampling distributions have mean values. The following box describes the key relationship between the mean values of MSTr and MSE.

When H_0 is true $(\mu_1 = \mu_2 = \cdots = \mu_k)$,

$$\mu_{\text{MSTr}} = \mu_{\text{MSE}}$$

However, when H_0 is false,

$$\mu_{\text{MSTr}} > \mu_{\text{MSE}}$$

and the greater the differences among the μ's, the larger μ_{MSTr} will be relative to μ_{MSE}.

According to this result, when H_0 is true, values of the two mean squares are expected to be close. However, MSTr will be substantially greater than MSE when some μ's differ greatly from others. This means that a calculated MSTr that is much greater than MSE is inconsistent with H_0.

In Example 15.2, MSTr = 10.0 and MSE = 0.70, so MSTr is more than 10 times as large as MSE. Can this difference be attributed solely to sampling variability, or is the ratio MSTr/MSE large enough to suggest that H_0 should be rejected? Before a formal

Copyright 2025 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s). Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

test procedure is introduced, it is necessary to revisit *F* distributions, which were first introduced in Chapter 14.

F distributions arise in connection with ratios of mean squares. A particular *F* distribution is obtained by specifying both numerator degrees of freedom (df_1) and denominator degrees of freedom (df_2).

Figure 15.4 shows an *F* curve for a particular choice of df_1 and df_2. All *F* tests in this text are upper-tailed, so *P*-values are areas under the *F* curve to the right of the calculated values of *F*.

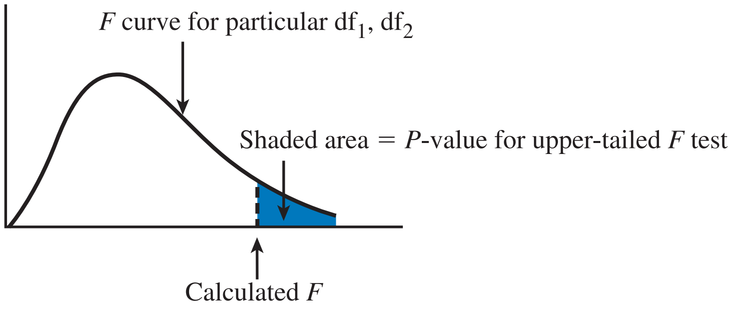

F curve for particular df_1, df_2
Shaded area = *P*-value for upper-tailed *F* test
Calculated *F*

Figure 15.4
An *F* curve and *P*-value for an upper-tailed test.

Many statistical software packages and graphing calculators can provide *P*-values for *F* tests. If you don't have access to technology, you will need to rely on tables. Tabulation of *F* curve upper-tail areas is cumbersome, because there are two degrees of freedom rather than just one (as in the case of *t* distributions). For selected (df_1, df_2) pairs, the *F* table (Appendix Table 6) gives only the four numbers that capture tail areas 0.10, 0.05, 0.01, and 0.001, respectively. For example, here are the four numbers for $\text{df}_1 = 4$, $\text{df}_2 = 10$ along with the statements that can be made about the *P*-value:

Tail area		0.10		0.05		0.01		0.001	
Value		2.61		3.48		5.99		11.28	
	↑		↑		↑		↑		↑
	a		b		c		d		e

a. $F < 2.61 \longrightarrow$ tail area $=$ *P*-value > 0.10
b. $2.61 < F < 3.48 \longrightarrow 0.05 < P\text{-value} < 0.10$
c. $3.48 < F < 5.99 \longrightarrow 0.01 < P\text{-value} < 0.05$
d. $5.99 < F < 11.28 \longrightarrow 0.001 < P\text{-value} < 0.01$
e. $F > 11.28 \longrightarrow P\text{-value} < 0.001$

For example, if $F = 7.12$, then $0.001 < P\text{-value} < 0.01$. If a test with $\alpha < 0.05$ is used, H_0 would be rejected, because $P\text{-value} \leq \alpha$.

Single-Factor ANOVA *F* Test

Null Hypothesis: $H_0\colon \mu_1 = \mu_2 = \cdots = \mu_k$

Test Statistic: $F = \dfrac{\text{MSTr}}{\text{MSE}}$

When H_0 is true and the ANOVA assumptions are reasonable, *F* has an *F* distribution with $\text{df}_1 = k - 1$ and $\text{df}_2 = N - k$.

The *P*-value is the area captured in the upper tail of the corresponding *F* curve. A statistical software package, a graphing calculator, or Appendix Table 6 can be used to determine *P*-values for *F* tests.

Example 15.3 Ball-to-Head Impact Revisited

The two mean squares for the peak linear acceleration data given in Example 15.1 were calculated in Example 15.2 to be

$$\text{MSTr} = 10.00 \qquad \text{MSE} = 0.700$$

The value of the F statistic is then

$$F = \frac{\text{MSTr}}{\text{MSE}} = \frac{10.00}{0.700} = 14.29$$

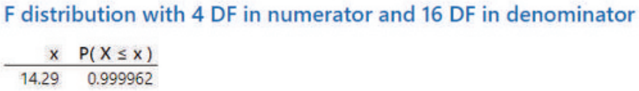
with $\text{df}_1 = k - 1 = 4$ and $\text{df}_2 = N - k = 20 - 4 = 16$.

Using $\text{df}_1 = 4$ and $\text{df}_2 = 16$, Appendix Table 6 shows that 7.94 captures tail area 0.001. Since $14.29 > 7.94$, it follows that P-value = (area to the right of 14.29) < 0.001.

Interpret the results ➤

The P-value is smaller than any reasonable α, so the null hypothesis is rejected. There is convincing evidence that the mean peak linear acceleration is not the same for all five ball inflation levels. Techniques for determining which means differ are introduced in Section 15.2.

Technology can also be used to calculate F-curve areas. For example, Minitab will calculate cumulative area (area to the left), as shown in the following output:



```
F distribution with 4 DF in numerator and 16 DF in denominator

    x   P( X ≤ x )
14.29     0.999962
```

Because the P-value for the ANOVA F test is an upper-tail area, it follows that the P-value for this test is

$$P\text{-value} = 1 - 0.999962 = 0.000038$$

This is consistent with the statement that the P-value is less than 0.001, and would lead to the same conclusion.

Summarizing an ANOVA

ANOVA calculations are often summarized in a table called an ANOVA table. To understand such a table, one more sum of squares must be defined.

Total sum of squares, denoted by SSTo, is given by

$$\text{SSTo} = \sum_{\text{all } N \text{ obs.}} (x - \bar{\bar{x}})^2$$

with associated $\text{df} = N - 1$.

The relationship between the three sums of squares SSTo, SSTr, and SSE is

$$\text{SSTo} = \text{SSTr} + \text{SSE}$$

which is called the fundamental identity for single-factor ANOVA.

The quantity SSTo, the sum of squared deviations about the grand mean, is a measure of total variability in the data set consisting of all k samples. The quantity SSE results from measuring variability separately within each sample and then combining these measures. Within-sample variability is present regardless of whether or not H_0 is true. The magnitude of SSTr, on the other hand, does depend on whether the null hypothesis is true or false. The more the μ's differ from one another, the larger SSTr will tend to be. This means that SSTr represents variability that can be explained by differences among the means. An informal paraphrase of the fundamental identity for single-factor ANOVA is

$$\text{total variability} = \text{explained variability} + \text{unexplained variability}$$

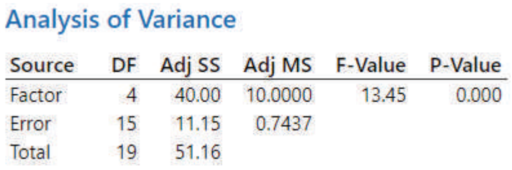
Once any two of the sums of squares have been calculated, the remaining one is easily obtained from the fundamental identity. Often SSTo and SSTr are calculated first (using technology or the computational formulas given in the online appendix to this chapter), and then SSE is obtained by subtraction: SSE = SSTo − SSTr. All the degrees of freedom, sums of squares, and mean squares are entered in an **ANOVA table**, as displayed in Table 15.2. The *P*-value usually appears to the right of *F* when the analysis is done using a statistical software package.

Table 15.2 General Format for a Single-Factor ANOVA Table

Source of Variation	df	Sum of Squares	Mean Square	*F*
Treatments	$k - 1$	SSTr	$\text{MSTr} = \dfrac{\text{SSTr}}{k-1}$	$F = \dfrac{\text{MSTr}}{\text{MSE}}$
Error	$N - k$	SSE	$\text{MSE} = \dfrac{\text{SSE}}{N-k}$	
Total	$N - 1$	SSTo		

An ANOVA table from Minitab for the peak linear acceleration data of Example 15.3 is shown in Table 15.3. The reported *P*-value is 0.000. The differences in the values of SSE, MSE, and *F* reported by Minitab and those found in Example 15.3 are due to rounding in the hand calculations.

Table 15.3 An ANOVA Table from Minitab for the Data of Example 15.3



```
Analysis of Variance

Source   DF   Adj SS    Adj MS   F-Value   P-Value
Factor    4    40.00   10.0000     13.45     0.000
Error    15    11.15    0.7437
Total    19    51.16
```

Example 15.4 Germs on Money and Credit Cards

Understand the context ➤

● Just how dirty are the money and the credit cards you carry in your wallet? The authors of the article **"Dirty Money: New Research Reveals the Filthiness of Our Cash, Cards, & Coins" (lendedu.com, 2020)** measured the bacteria on the surface of credit cards, bills, and coins. Germ scores (with higher scores corresponding to dirtier surfaces) for a sample of 21 bills of various denominations ranging from \$1 to \$100, a sample of 8 coins, and a sample of 27 credit cards are given here. (This is a subset of the data included in the paper.)

Germ Score for 21 Bills

32	35	39	41	54	55	56	68	74	78
98	104	105	108	119	119	126	129	143	163
200									

Germ Score for 8 Coins

25	32	33	52	60	84	92	109

Germ Score for 27 Credit Cards

48	63	68	75	104	106	106	112	121	122
124	133	134	134	139	139	143	143	146	149
150	156	160	163	168	172	179			

● **Data set available online**

Copyright 2025 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s). Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

So what do these germ scores mean? For reference, it is recommended that a food establishment surface should have a germ score of 10 or less. The average germ scores for other surfaces studied were 68 for a New York City subway pole, 163 for bathrooms at Penn Station in New York City, and 664 for door handles at McDonald's!

Summary statistics for the bills, coins, and credit card data follow.

	Bills	**Coins**	**Credit Cards**
n	21	8	27
$\bar{x}$	92.67	60.90	128.04
s	45.47	31.20	35.05

Also, $N = 56$, the grand total $= 5890$, and $\bar{\bar{x}} = \dfrac{5890}{56} = 105.18$

Formulate a plan ➤

You can carry out an F test to see if there is convincing evidence that the mean germ score is not the same for bills, coins, and credit cards.

1. Use μ_B, μ_C, and μ_{CC} to denote the population mean germ score for bills, coins, and credit cards, respectively.
2. H_0: $\mu_B = \mu_C = \mu_{CC}$
3. H_a: at least two of the population means differ
4. Significance level: $\alpha = 0.05$
5. Test statistic: $F = \dfrac{\text{MSTr}}{\text{MSE}}$
6. Assumptions: Figure 15.5 shows boxplots of the data from each of the three samples. The boxplots are approximately symmetric and there are no outliers. The largest standard deviation ($s_B = 45.47$) is not more than twice as large as the smallest ($s_C = 31.20$). If it is reasonable to assume that the three samples are representative of the populations of bills, coins, and credit cards, then the assumptions of ANOVA are reasonable.

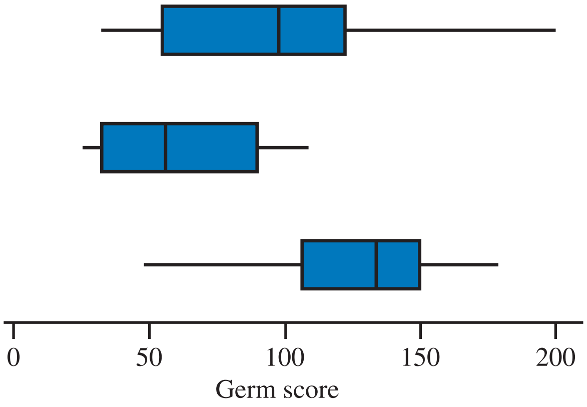
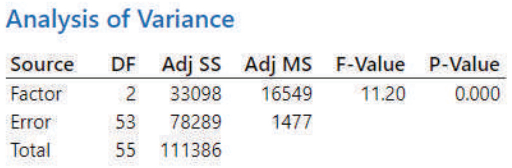


Figure 15.5
Boxplots for the data of Example 15.4.

Do the work ➤

7. Calculation:
From the accompanying Minitab output, $F = 11.20$



```
Analysis of Variance

Source  DF  Adj SS  Adj MS  F-Value  P-Value
Factor   2   33098   16549    11.20    0.000
Error   53   78289    1477
Total   55  111386
```

8. P-value: From the given Minitab output, P-value $= 0.000$.

Interpret the results ➤

9. Conclusion: The P-value of 0.000 is less than the selected significance level, so you reject the null hypothesis. There is convincing evidence that the mean germ scores for bills, coins, and credit cards are not all equal.

Copyright 2025 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s). Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

Exercises 15.1–15.13

● **Data set available online**

15.1 Using technology or Appendix Table 6, give as much information as you can about the P-value for an upper-tailed F test in each of the following situations:

a. $df_1 = 4, df_2 = 15, F = 5.37$
b. $df_1 = 4, df_2 = 15, F = 1.90$
c. $df_1 = 4, df_2 = 15, F = 4.89$
d. $df_1 = 3, df_2 = 20, F = 14.48$
e. $df_1 = 3, df_2 = 20, F = 2.69$
f. $df_1 = 4, df_2 = 50, F = 3.24$

15.2 Using technology or Appendix Table 6, give as much information as you can about the P-value of the single-factor ANOVA F test in each of the following situations:

a. $k = 5, n_1 = n_2 = n_3 = n_4 = n_5 = 4, F = 5.37$
b. $k = 5, n_1 = n_2 = n_3 = 5, n_4 = n_5 = 4, F = 2.83$
c. $k = 3, n_1 = 4, n_2 = 5, n_3 = 6, F = 5.02$
d. $k = 3, n_1 = n_2 = 4, n_3 = 6, F = 15.90$
e. $k = 4, n_1 = n_2 = 15, n_3 = 12, n_4 = 10, F = 1.75$

15.3 Employees of a state university system can choose from among four different health plans. Each plan differs somewhat from the others in terms of hospitalization coverage. Four random samples of recently hospitalized individuals were selected, each sample consisting of people covered by a different health plan. The length of the hospital stay (number of days) was determined for each individual selected.

a. What hypotheses would you test to decide whether the mean lengths of stay are not the same for all four health plans? (Note: Carefully define the population characteristics of interest.)
b. If each sample consisted of eight individuals and the value of the ANOVA F statistic was $F = 4.37$, what conclusion would be appropriate for a test with $\alpha = 0.01$? (Hint: See Example 15.3.)
c. Answer the question posed in Part (b) if the F value given there resulted from sample sizes $n_1 = 9, n_2 = 8, n_3 = 7$, and $n_4 = 8$.

15.4 The paper **"Human-Induced Morphological Shifts in an Island Lizard" (*Evolutionary Applications* [2011]: 388–396)** describes a study of the effect of habitat on the brown anole lizard, a species that can be found in both natural and human-disturbed habitats. Data on weight (in grams) from representative samples of lizards from natural habitats on each of four islands in the Bahamas (Bimini, Great Exuma, Harbor, and New Providence) were used to produce the summary statistics in the accompanying table. Do these data provide convincing evidence that the mean weight for this species of lizard is not the same for all four island lizard populations? Test the relevant hypotheses using a significance level of α = 0.01.

Island	Sample Mean Weight	Sample Standard Deviation	Sample Size
Bimini	3.73	0.55	20
Great Exuma	4.48	0.88	20
Harbor	4.35	0.88	20
New Providence	3.17	0.66	21

$N = 81$, grand total $T = 319.4$, and $\bar{\bar{x}} = 3.94$

15.5 ● The authors of the paper **"Age and Violent Content Labels Make Video Games Forbidden Fruits for Youth" (*Pediatrics* [2009]: 870–876)** carried out an experiment to determine if restrictive labels on video games actually increased the attractiveness of the game for young game players. Participants read a description of a new video game and were asked how much they wanted to play the game. The description also included an age rating. Some participants read the description with an age restrictive label of 7+, indicating that the game was not appropriate for children under the age of 7. Others read the same description, but with an age restrictive label of 12+, 16+, or 18+. The following data for 7- to 8-year-old children are consistent with summary statistics given in the paper. (The sample sizes in the actual experiment were larger.) For purposes of this exercise, you can assume that the children were assigned at random to one of the four age label treatments (7+, 12+, 16+, and 18+). Data shown are the children's ratings of how much they wanted to play the game on a scale of 1 to 10. Do the data provide convincing evidence that the mean rating associated with the game description by 7- to 8-year-old children is not the same for all four restrictive rating labels? Test the appropriate hypotheses using a significance level of 0.05.

7+ label	12+ label	16+ label	18+ label
8	6	9	10
4	6	9	6
6	8	6	8
5	5	8	7
6	8	10	9
6	7	8	8
6	8	6	10
10	8	7	10
5	5	11	8
7	8	3	9

15.6 ● The previous exercise also gave data for 12- to 13-year-old children. Data consistent with summary values in the paper are shown on the next page. Do the data provide convincing evidence that the mean rating associated with the game description for 12- to 13-year-old children is not the same for all four age restrictive rating labels? Test the appropriate hypotheses using $\alpha = 0.05$.

7+ label	**12+ label**	**16+ label**	**18+ label**
4	6	9	9
7	5	7	7
5	6	8	4
9	5	6	9
4	7	8	7
3	6	10	10
6	8	9	8
3	7	6	8
6	8	8	8
5	6	7	7

15.7 ● Do people feel hungrier after sampling a healthy food? The authors of the paper **"When Healthy Food Makes You Hungry" (*Journal of Consumer Research* [2010]: S34–S44)** carried out a study to answer this question. They randomly assigned volunteers into one of three groups. The people in the first group were asked to taste a snack that was described as a new health bar containing high levels of protein, vitamins, and fiber. The people in the second group were asked to taste the same snack, but were told it was a tasty chocolate bar with a raspberry center. After tasting the snack, participants were asked to rate their hunger level on a scale from 1 (not at all hungry) to 7 (very hungry). The people in the third group were asked to rate their hunger but were not given a snack.

The data in the accompanying table are consistent with summary quantities given in the paper (although the sample sizes in the actual study were larger).

a. Do these data provide evidence that the mean hunger rating is not the same for all three treatments ("healthy" snack, "tasty" snack, no snack)? Test the relevant hypotheses using a significance level of 0.05.

b. Is it reasonable to conclude that the mean hunger rating is greater for people who do not get a snack? Explain.

Treatment Group	**Hunger Rating**									**Sample Mean**	**Sample Standard Deviation**
Healthy	4	7	7	4	5	3	4	7	6	5.2	1.56
Tasty	4	1	3	2	6	2	5	3	4	3.3	1.58
No Snack	3	4	5	6	5	4	2	4	4	4.1	1.17
							Overall mean = 4.2				

15.8 The authors of the paper **"Reading Subtitles and Taking Enotes While Learning Scientific Materials in a Multimedia Environment" (*Educational Technology and Society* [2016]: 47–58)** were interested in determining if including subtitles and providing opportunities to take electronic notes while listening to online materials would enhance learning for students whose first language is not English. Students were randomly assigned to one of four groups. In the first group, subtitles were included in the online materials that the students were asked to study, but the ability to take electronic notes was not provided. For the second group, no subtitles were provided, but students were able to take electronic notes. For the third group, both subtitles and the ability to take electronic notes were available. Students in the fourth group did not have access to either subtitles or the ability to take electronic notes. After studying the online materials, all students took a 14-question test on the material studied.

Minitab output based on data consistent with summary quantities in the paper is shown below. Is there convincing evidence that the mean test score is not the same for the four different treatments? Use the given computer output to test the appropriate hypotheses with a significance level of 0.05.

```
Analysis of Variance

Source    DF    Adj SS     Adj MS    F-Value    P-Value
Factor     3     37.76     12.558     2.39       0.077
Error     63    332.36      5.276
Total     66    370.13
```

15.9 The study described in Exercise 15.4 also gave data on body mass (in grams) from representative samples of lizards from disturbed habitats on each of four islands in the Bahamas (Bimini, Great Exuma, Harbor, and New Providence). Summary statistics are given in the accompanying table. Do these data provide convincing evidence that the mean mass for this species of lizards is not the same for all four island lizard populations? You can assume that the mass distribution is approximately normal for each island.

Island	**Sample Mean Weight**	**Sample Standard Deviation**	**Sample Size**
Bimini	4.22	0.77	20
Great Exuma	7.11	1.41	18
Harbor	5.75	1.10	20
New Providence	5.13	1.22	20

$N = 78$, grand total $T = 430.00$, and $\bar{\bar{x}} = 5.51$

Copyright 2025 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s). Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

15.10 The paper **"Effects of Mindfulness Training and Exercise on Cognitive Function in Older Adults" (*Journal of the American Medical Association* [2022]: 2218–2229)** describes and experiment to investigate whether meditation and exercise could improve memory and executive function. Study participants were 585 older adults with reduced cognitive functioning. Participants were randomly assigned to one of four groups. Those in group 1 meditated for about 60 minutes per day, those in group 2 performed 300 minutes of exercise per week, those in group 3 did both meditation and exercise, and those in group 4 were a control group that did neither meditation nor exercise. The following conclusion is from the paper:

> Among older adults with subjective cognitive concerns, mindfulness training, exercise, or both did not result in significant improvement in episodic memory or executive function at 6 months. The findings do not support the use of these interventions for improving cognition in older adults with subjective cognitive concerns.

a. The conclusion above is based on two hypothesis tests, one considering the response on measure of memory and one considering the response on a measure of executive function. What hypotheses were tested for the conclusion about executive function?

b. Did the researchers reject the null hypothesis stated in Part (a) or fail to reject the null hypothesis?

c. Explain how the statement "The findings do not support the use of these interventions for improving cognition in older adults with subjective cognitive concerns" is supported by the outcomes of the hypothesis tests.

15.11 In the experiment described in the previous exercise, $N = 585$ and the number of treatments was $k = 4$. To reach the conclusion that there was no significant improvement in episodic memory, the researchers might have tested the null hypothesis that there was no difference in the mean improvement on the measure of episodic memory.

a. State the null and alternative hypotheses that might have been tested.

b. What are the degrees of freedom for the ANOVA F test of these hypotheses?

c. Given the stated conclusion, which of the following might have been the value of the F statistic?

1.5
11.5
111.5

15.12 It is common for baseball pitchers to use stretching to prepare for a game. But does this make a difference? The authors of the paper **"The Acute Effects of Upper Extremity Stretching on Throwing Velocity in Baseball Throwers" (*Journal of Sports Medicine* [2013]: 1–7)** carried out an experiment to compare two different types of stretching and a control treatment consisting of no stretching. Participants had varying levels of baseball throwing experience and were not professional or collegiate baseball players. Participants in the two stretching treatments went through a warm-up that included 8 minutes of stretching. Each participant then threw 10 pitches, and the average speed (km/hour) was calculated.

a. Explain why it is important that the participants be assigned at random to the three different treatment groups (Stretching Method 1, Stretching Method 2, and No Stretching).

b. The following computer output and summary values are based on simulated data that are consistent with information and conclusions given in the paper. Use the given output to determine if there is evidence to support the claim that the means for average speed are not the same for all three treatments. Use a significance level of 0.05 for your test.

```
Analysis of Variance

Source     DF    Adj SS    Adj MS    F-Value    P-Value
Factor      2     1097     548.7      2.44       0.094
Error      78    17570     225.3
Total      80    18667

Means

Factor              N     Mean    StDev          95% CI
Stretching         27    86.26    15.95      (80.51,92.01)
  Method 1
Stretching         27    90.30    17.06      (84.55,96.05)
  Method 2
No Stretching      27    95.26    11.42      (89.51,101.01)
```

c. Previous research on the effect of stretching on performance in other sports, such as running, has concluded that stretching can improve performance. Why do you think that the authors of this paper were surprised by the results of this study?

15.13 In an experiment to investigate the performance of four different brands of spark plugs intended for use on a 125-cc motorcycle, five plugs of each brand were tested, and the number of miles (at a constant speed) until failure was observed. A partially completed ANOVA table is given. Fill in the missing entries, and test the relevant hypotheses using a 0.05 level of significance. (Hint: See Table 15.2.)

Source of Variation	df	Sum of Squares	Mean Square	F
Treatments				
Error		235,419.04		
Total		310,500.76		

Copyright 2025 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s). Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

Section 15.2 Multiple Comparisons

When H_0: $\mu_1 = \mu_2 = \cdots = \mu_k$ is rejected by the F test, there is convincing evidence that there are differences among the k population or treatment means. A natural question to ask then is, Which means differ? For example, with $k = 4$, it might be the case that $\mu_1 = \mu_2 = \mu_4$, with μ_3 different from the other three means. Another possibility is that $\mu_1 = \mu_4$ and $\mu_2 = \mu_3$. Still another possibility is that all four means are different from one another.

A **multiple comparisons procedure** is a method for identifying differences among the μ's once the hypothesis that all the means are equal has been rejected. One such method is the **Tukey-Kramer** (T-K) multiple comparisons procedure.

The T-K procedure is based on calculating confidence intervals for the difference between each possible pair of μ's. For example, for $k = 3$, there are three differences to consider:

$$\mu_1 - \mu_2 \qquad \mu_1 - \mu_3 \qquad \mu_2 - \mu_3$$

(The difference $\mu_2 - \mu_1$ is not considered, because the interval for $\mu_1 - \mu_2$ provides the same information. Similarly, intervals for $\mu_3 - \mu_1$ and $\mu_3 - \mu_2$ are not necessary.)

After all confidence intervals have been calculated, each is examined to determine whether the interval includes 0. If a particular interval does not include 0, the two means are declared "significantly different" from one another. An interval that does include 0 supports the conclusion that there may be no difference between the means involved.

Suppose, for example, that $k = 3$ and that the three confidence intervals are

Difference	T-K Confidence Interval
$\mu_1 - \mu_2$	$(-0.9, 3.5)$
$\mu_1 - \mu_3$	$(2.6, 7.0)$
$\mu_2 - \mu_3$	$(1.2, 5.7)$

Because the interval for $\mu_1 - \mu_2$ includes 0, you would say that μ_1 and μ_2 do not differ significantly. The other two intervals do not include 0, so you would conclude that $\mu_1 \neq \mu_3$ and $\mu_2 \neq \mu_3$.

The T-K intervals are based on critical values for a probability distribution called the *Studentized range distribution*. These critical values appear in Appendix Table 7. To find a critical value, enter the table at the column corresponding to the number of populations or treatments being compared, move down to the row corresponding to the number of error degrees of freedom, and select either the value for a 95% confidence level or the one for a 99% level.

The Tukey-Kramer (T-K) Multiple Comparison Procedure

When there are k populations or treatments being compared, $\dfrac{k(k-1)}{2}$ confidence intervals are calculated.

Denoting the relevant Studentized range critical value (from Appendix Table 7) by q, the intervals are as follows:

$$\text{For } \mu_i - \mu_j\text{:} \quad (\bar{x}_i - \bar{x}_j) \pm q\sqrt{\frac{\text{MSE}}{2}\left(\frac{1}{n_i} + \frac{1}{n_j}\right)}$$

Two means are judged to be significantly different if the corresponding interval does not include zero.

Copyright 2025 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s). Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

If the sample sizes are all equal, n can be used to denote the common value of $n_1, \dots, n_k$. In this case, the ± term for all of the intervals is the same quantity

$$q\sqrt{\frac{MSE}{n}}$$

Example 15.5 Dirty Money Revisited

Understand the context ➤ Example 15.4 gave data and summary statistics on germ scores measured on the surface of bills, coins, and credit cards. You can use μ_B, μ_C, and μ_{CC} to denote the population mean germ score for bills, coins, and credit cards, respectively.

From Example 15.4, MSTr = 16,549, MSSE = 1,477, $F = 11.20$, and P-value = 0.000. This led to the conclusion that there is convincing evidence that the mean germ scores for bills, coins, and credit cards are not all equal.

Summary statistics for the bills, coins, and credit card data were given in Example 15.4 and are also shown here:

	Bills	**Coins**	**Credit Cards**
n	21	8	27
$\bar{x}$	92.67	60.90	128.04
s	45.47	31.20	35.05

Do the work ➤ Appendix Table 7 gives the 95% studentized range critical value $q = 3.40$ (using $k = 3$ and error df = 60, the closest value to df $= N - k = 53$).

The T-K intervals are

$$\mu_B - \mu_C\text{: } (92.67 - 60.90) \pm 3.40\sqrt{\left(\frac{1477}{2}\right)\left(\frac{1}{21} + \frac{1}{8}\right)}$$

$$= 31.77 \pm 38.39$$

$$= (-6.62, 70.16)$$

Includes 0

$$\mu_B - \mu_{cc}\text{: } (92.67 - 128.40) \pm 3.40\sqrt{\left(\frac{1477}{2}\right)\left(\frac{1}{21} + \frac{1}{27}\right)}$$

$$= -35.37 \pm 26.88$$

$$= (-62.25, -8.49)$$

Does not Includes 0

$$\mu_C - \mu_{CC}\text{: } (60.90 - 128.40) \pm 3.40\sqrt{\left(\frac{1477}{2}\right)\left(\frac{1}{8} + \frac{1}{27}\right)}$$

$$= -67.14 \pm 37.19$$

$$= (-104.33, -29.95)$$

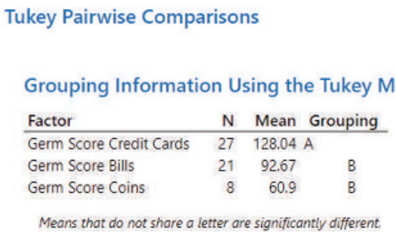
Does not Includes 0

Interpret the results ➤ Because the interval comparing bills and coins includes 0, you would conclude that the mean germ score for bills is not significantly different from mean germ score for coins. The intervals comparing bills to credit cards and coins to credit cards do not include 0, so you would conclude that the mean germ score for credit cards is significantly different from the mean germ scores of bills and coins.

Minitab and other statistical software packages can also be used to construct T-K intervals if raw data is available. Figure 15.6 shows the Minitab summary for the germ score data of Example 15.5. Notice that Minitab uses letters rather than line segments to indicate significant differences. Bills and coins share the same letter in the display, indicating that the mean germ scores for these two populations were not significantly different. Credit

Copyright 2025 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s). Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

cards have a letter that is different, indicating that the mean germ score for credit cards is significantly different from the other two.



Tukey Pairwise Comparisons

Grouping Information Using the Tukey Method and 95% Confidence

Factor	N	Mean	Grouping
Germ Score Credit Cards	27	128.04	A
Germ Score Bills	21	92.67	B
Germ Score Coins	8	60.9	B

Means that do not share a letter are significantly different.

Figure 15.6 Minitab T-K summary for Example 15.5.

Why calculate the T-K intervals rather than use the t confidence interval for a difference between two means introduced in Chapter 11? The answer is that the T-K intervals control the **simultaneous confidence level** at approximately 95% (or 99%). That is, if the procedure is used repeatedly on many different data sets, in the long run only about 5% (or 1%) of the time would *at least one* of the intervals not include the value that the interval is estimating. For example, consider using separate 95% t intervals, each one having a 5% error rate. When this is done, the chance that at least one interval would make an incorrect statement about a difference in μ's increases dramatically with the number of intervals calculated. For example, to achieve a simultaneous confidence level of about 95% (experimentwise or "family" error rate of 5%) when $k = 4$ and error df $= 76$, the confidence level for individual intervals must be 98.95% (individual error rate 1.05%).

An effective display for summarizing the results of any multiple comparisons procedure involves listing the $\bar{x}$'s and underscoring pairs judged to be not significantly different. The process for constructing such a display is described in the box below.

To illustrate this summary procedure, suppose that four samples with $\bar{x}_1 = 19$, $\bar{x}_2 = 27$, $\bar{x}_3 = 24$, and $\bar{x}_4 = 10$ are used to test $H_0: \mu_1 = \mu_2 = \mu_3 = \mu_4$ and that this hypothesis is rejected. Suppose the T-K confidence intervals indicate that μ_2 is significantly different from both μ_1 and μ_4, and that there are no other significant differences. The resulting summary display would then be

Population	4	1	3	2
Sample mean	10	19	24	27

Summarizing the Results of the Tukey-Kramer (T-K) Procedure

1. List the sample means in increasing order, identifying the corresponding population just above the value of each $\bar{x}$.
2. Use the T-K intervals to determine the group of means that do not differ significantly from the first mean in the list. Draw a horizontal line extending from the smallest mean to the last mean in the group identified. For example, if there are five means, arranged in order,

Population	3	2	1	4	5
Sample mean	$\bar{x}_3$	$\bar{x}_2$	$\bar{x}_1$	$\bar{x}_4$	$\bar{x}_5$

and μ_3 is judged to be not significantly different from μ_2 or μ_1, but is judged to be significantly different from μ_4 and μ_5, draw the following line:

Population	3	2	1	4	5
Sample mean	$\bar{x}_3$	$\bar{x}_2$	$\bar{x}_1$	$\bar{x}_4$	$\bar{x}_5$

3. Use the T-K intervals to determine the group of means that are not significantly different from the second smallest mean. (You need consider only means that appear to the right of the mean under consideration.) If there is already a line connecting the second smallest mean with all means in the new group identified,

(continued)

Copyright 2025 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s). Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

no new line need be drawn. If this entire group of means is not underscored with a single line, draw a line extending from the second smallest mean to the last mean in the new group. Continuing with the example, if μ_2 is not significantly different from μ_1 but is significantly different from μ_4 and μ_5, no new line needs to be drawn. However, if μ_2 is not significantly different from either μ_1 or μ_4 but is judged to be different from μ_5, a second line is drawn as shown:

Population	3	2	1	4	5
Sample mean	$\bar{x}_3$	$\bar{x}_2$	$\bar{x}_1$	$\bar{x}_4$	$\bar{x}_5$

4. Continue considering the means in the order listed, adding new lines as needed.

Example 15.6 Sleep Time

Understand the context ➤

● Suppose that a biologist studied the effects of ethanol on sleep time. A sample of 20 rats of the same age was selected, and rats were randomly assigned to one of four treatment groups. Rats were given an oral injection having a particular concentration of ethanol per ounce of body weight. The treatments were different amounts of ethanol. The rapid eye movement (REM) sleep time for each rat was then recorded for a 24-hour period, with the results shown in the following table:

Treatment	Observations					$\bar{x}$
1. 0 (control)	88.6	73.2	91.4	68.0	75.2	79.28
2. 1 g/kg	63.0	53.9	69.2	50.1	71.5	61.54
3. 2 g/kg	44.9	59.5	40.2	56.3	38.7	47.92
4. 4 g/kg	31.0	39.6	45.3	25.2	22.7	32.76

Table 15.4 (an ANOVA table from SAS) leads to the conclusion that actual mean REM sleep time is not the same for all four treatments; the *P*-value for the *F* test is 0.0001.

Table 15.4 **SAS ANOVA Table for Example 15.6**

```
Analysis of Variance Procedure Dependent Variable: TIME
                          Sum of           Mean
Source         DF        Squares         Square      F Value      Pr > F
Model           3     5882.35750     1960.78583        21.09      0.0001
Error          16     1487.40000       92.96250
Total          19     7369.75750
```

Do the work ➤

The T-K intervals are

Difference	Interval	Includes 0?
$\mu_1 - \mu_2$	17.74 ± 17.446	no
$\mu_1 - \mu_3$	31.36 ± 17.446	no
$\mu_1 - \mu_4$	46.24 ± 17.446	no
$\mu_2 - \mu_3$	13.08 ± 17.446	yes
$\mu_2 - \mu_4$	28.78 ± 17.446	no
$\mu_3 - \mu_4$	15.16 ± 17.446	yes

Interpret the results ➤

The only T-K intervals that include zero are those for $\mu_2 - \mu_3$ and $\mu_3 - \mu_4$. The corresponding underscoring pattern is

$\bar{x}_4$	$\bar{x}_3$	$\bar{x}_2$	$\bar{x}_1$
32.76	47.92	61.54	79.28

Figure 15.7 displays the SAS output that agrees with this underscoring. In the SAS output, letters are used to indicate groupings in place of the underscoring.

● **Data set available online**

Copyright 2025 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s). Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

Figure 15.7
SAS output for Example 15.6.

```
Alpha = 0.05 df = 16 MSE = 92.9625
Critical Value of Studentized Range = 4.046
Minimum Significant Difference = 17.446
Means with the same letter are not significantly different.
Tukey Grouping     Mean     N     Treatment
       A         79.280     5   0 (control)
       B         61.540     5        1 g/kg
C      B         47.920     5        2 g/kg
C                32.760     5        4 g/kg
```

Example 15.7 Mug Color and the Taste of Coffee

The paper **"Does the Colour of the Mug Influence the Taste of the Coffee?" (*Flavour* [2014]: 1–7)** describes an experiment to investigate whether the color of a coffee mug has an effect on how people perceive the flavor intensity of coffee. In this experiment, 18 volunteers were randomly assigned to three groups. All were served the same coffee, but those in first group received their coffee in a white mug, those in the second group received their coffee in a blue mug, and those in the third group received their coffee in a clear glass mug. After tasting the coffee, each person rated the flavor intensity of the coffee. The means for intensity rating in the three treatment groups (approximate values from a graph in the paper) were 50 for the white mug group, 42 for the blue mug group, and 23 for the clear glass mug group.

The researchers used an ANOVA *F* test to test the null hypothesis of no difference in the mean intensity rating for the three treatments. They reported an *F* statistic value of 4.78 and a *P*-value of 0.025. Based on this *P*-value, they rejected the null hypothesis and concluded that the means for intensity rating were not equal for all three mug colors. A multiple comparison procedure was used to identify differences in treatment means, and these conclusions were reached:

- Coffee was rated as significantly more intense when served in a white mug than when served in a clear glass mug.
- None of the other differences were found to be significant.

These results are summarized in the following underscoring pattern. The treatment group means are arranged in order from smallest to largest.

Clear	Blue	White
23	42	50

Based on this analysis, you would conclude that there is a difference between the means for intensity rating for coffee served in a white mug and coffee served in a clear glass mug. The effect of a blue mug is unclear, and there is not convincing evidence of a difference between the means for intensity rating of coffee served in a blue mug and coffee served in a white mug. There is also no convincing evidence of a difference for coffee served in a blue mug and coffee served in a clear glass mug.

Exercises 15.14–15.22

● **Data set available online**

15.14 Leaf surface area is an important variable in plant gas-exchange rates. Dry matter per unit surface area (mg/cm$_3$) was measured for trees raised under three different growing conditions. Let μ_1, μ_2, and μ_3 represent the mean dry matter per unit surface area for the growing conditions 1, 2, and 3, respectively. Suppose that the 95% T-K confidence intervals are

Difference	**Interval**
$\mu_1 - \mu_2$:	$(-3.11, -1.11)$
$\mu_1 - \mu_3$:	$(-4.06, -2.06)$
$\mu_2 - \mu_3$:	$(-1.95, 0.05)$

Which of the following four statements do you think describes the relationship between μ_1, μ_2, and μ_3? Explain your choice. (Hint: See Example 15.5.)

a. $\mu_1 = \mu_2$, and μ_3 differs from μ_1 and μ_2.
b. $\mu_1 = \mu_3$, and μ_2 differs from μ_1 and μ_3.
c. $\mu_2 = \mu_3$, and μ_1 differs from μ_2 and μ_3.
d. All three μ's are different from one another.

15.15 The paper **"Trends in Blood Lead Levels and Blood Lead Testing among U.S. Children Aged 1 to 5 Years" (*Pediatrics* [2009]: e376–e385)** gave data on blood lead levels (in μg/dL) for samples of children living in homes that had been classified either at low, medium, or high risk of lead exposure based on when the home was constructed. After using a multiple comparison procedure, the authors reported the following:

1. The difference in mean blood lead level between low-risk housing and medium-risk housing was significant.
2. The difference in mean blood lead level between low-risk housing and high-risk housing was significant.
3. The difference in mean blood lead level between medium-risk housing and high-risk housing was significant.

Which of the following sets of T-K intervals (Set 1, 2, or 3) is consistent with the authors' conclusions? Explain your choice. (Hint: See Example 15.5.)

μ_L = mean blood lead level for children living in low-risk housing
μ_M = mean blood lead level for children living in medium-risk housing
μ_H = mean blood lead level for children living in high-risk housing

Difference	Set 1	Set 2	Set 3
$\mu_L - \mu_M$	$(-0.6, 0.1)$	$(-0.6, -0.1)$	$(-0.6, -0.1)$
$\mu_L - \mu_H$	$(-1.5, -0.6)$	$(-1.5, -0.6)$	$(-1.5, -0.6)$
$\mu_M - \mu_H$	$(-0.9, -0.3)$	$(-0.9, 0.3)$	$(-0.9, -0.3)$

15.16 The paper **"Effect of Cooking Methods on Glycemic Index and In Vitro Bioaccessibility of Potato (*Solanum tuberosum L.*) Carbohydrates" (*LWT—Food Science and Technology* [2020] 109363)** describes a study investigating the effect of different cooking methods on the starch and sugar content of potatoes. The starch and sugar content were measured for raw potatoes and for potatoes that were cooked using three different cooking methods (boiling, frying, and microwaving). After concluding that the mean starch content was not the same for the four treatments (raw, boiled, fried, and microwaved), T-K intervals were used identify which treatment means were significantly different. The accompanying underscoring pattern is based on the T-K intervals. Write a few sentences interpreting this underscoring pattern. (Hint: See Example 15.7.)

$$\underline{\text{Raw}\qquad\text{Fried}}\qquad\underline{\text{Boiled}}\qquad\underline{\text{Microwaved}}$$

15.17 The paper referenced in the previous exercise also recorded sugar content for potatoes in the four treatments.

$$\begin{array}{cccc} \text{Raw} & \text{Boiled} & \text{Microwaved} & \text{Fried} \\ \cline{1-2} \\[-2ex] \cline{2-3} \\[-2ex] \cline{4-4} \end{array}$$

a. Write a few sentences interpreting this underscoring pattern.
b. Using your answers from Part (a) and from the previous exercise, write a few sentences describing the differences that the cooking methods have on starch compared to sugar.

15.18 ● The paper referenced in Exercise 15.5 described an experiment to determine if restrictive age labeling on video games increased the attractiveness of the game for 7- to 8-year-old children. In that exercise, the null hypothesis was $H_0: \mu_1 = \mu_2 = \mu_3 = \mu_4$, where μ_1 is the population mean attractiveness rating for the game with the 7+ age label, and μ_2, μ_3, and μ_4 are the population mean attractiveness scores for the 12+, 16+, and 18+ age labels, respectively. The sample data are given in the accompanying table.

7+ label	12+ label	16+ label	18+ label
8	6	9	10
4	6	9	6
6	8	6	8
5	5	8	7
6	8	10	9
6	7	8	8
6	8	6	10
10	8	7	10
5	5	11	8
7	8	3	9

a. Calculate the 95% T-K intervals and then use the underscoring procedure described in this section to identify significant differences among the age labels.
b. Based on your answer to Part (a), write a few sentences commenting on the theory that the more restrictive the age label on a video game, the more attractive the game is to 7- to 8-year-old children.

15.19 The study described in Exercise 15.9 **("Human-Induced Morphological Shifts in an Island Lizard," *Evolutionary Applications* [2011]: 388–396)** gave data on body mass (in grams) from representative samples of lizards from disturbed habitats on each of four islands in the Bahamas (Bimini, Great Exuma, Harbor, and New Providence). Summary statistics are given in the accompanying table.

An ANOVA F test leads to the conclusion that there is convincing evidence that the mean mass for this species of lizards is not the same for all four island lizard populations.

Island	Sample Mean Weight	Sample Standard Deviation	Sample Size
Bimini	4.22	0.77	20
Great Exuma	7.11	1.41	18
Harbor	5.75	1.10	20
New Providence	5.13	1.22	20
$N = 78$, grand total $T = 430.00$, and $\bar{\bar{x}} = 5.51$			

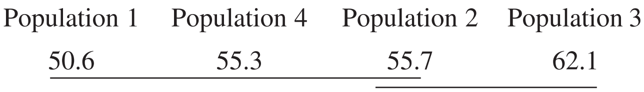
a. Use the T-K procedure to find the T-K intervals.

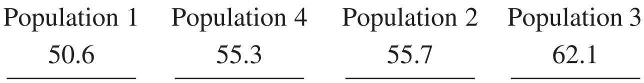
b. Use the intervals from Part (a) to construct the corresponding underscoring pattern and interpret the pattern in context.

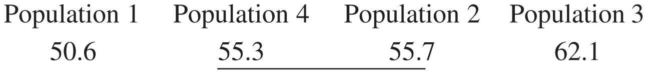
15.20 Data from random samples from each of four populations were used to test the null hypothesis that the four population means were all equal. This hypothesis was rejected, indicating that at least two of the population means are different. The Tukey-Kramer (T-K) confidence intervals were then calculated to be as shown in the accompanying table.

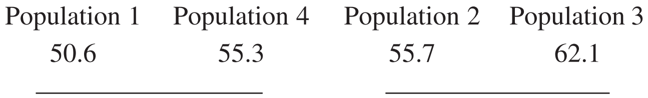
Difference	T-K Interval
Population 1 − Population 2	(−9.7, −0.5)
Population 1 − Population 3	(−16.2, −7.0)
Population 1 − Population 4	(−9.3. −0.1)
Population 2 − Population 3	(−11.1, −1.9)
Population 2 − Population 4	(−4.2, 5.0)
Population 3 − Population 4	(2.3, 11.5)

Which of the following underscoring patterns is consistent with the given T-K intervals?

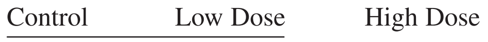








15.21 The paper **"Inflammatory and Oxidative Responses Induced by Exposure to Commonly Used e-Cigarette Flavoring Chemicals and Flavored e-Liquids without Nicotine" (*Frontiers in Physiology* [2018]: 1130)** describes a series of experiments to assess the effect of chemicals used as flavoring in e-cigarettes on two types of human cells. For one of the flavoring chemicals, diacetyl, the inflammatory response was measured for three samples of cells. One sample was a control and was not exposed to any flavoring chemical, one sample was exposed to a low dose of the flavoring chemical, and one sample was exposed to a high dose of the flavoring chemical. The accompanying underscoring pattern is based on T-K intervals reported in the paper.



Write a few sentences interpreting this underscoring.

15.22 ● Samples of six different brands of diet or imitation margarine were analyzed to determine the level of physiologically active polyunsaturated fatty acids (PAPUFA, in percent), resulting in the data shown in the accompanying table. (The data are fictitious, but the sample means agree with data reported in ***Consumer Reports***.)

Imperial	14.1	13.6	14.4	14.3	
Parkay	12.8	12.5	13.4	13.0	12.3
Blue Bonnet	13.5	13.4	14.1	14.3	
Chiffon	13.2	12.7	12.6	13.9	
Mazola	16.8	17.2	16.4	17.3	18.0
Fleischmann's	18.1	17.2	18.7	18.4	

a. Test for differences among the true means for PAPUFA percentage among the different brands. Use $\alpha = 0.05$.

b. Use the T-K procedure to calculate 95% simultaneous confidence intervals for all differences between means and give the corresponding underscoring pattern.

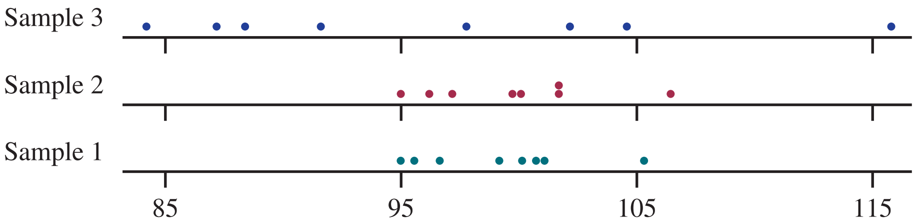
Copyright 2025 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s). Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

Chapter Activity

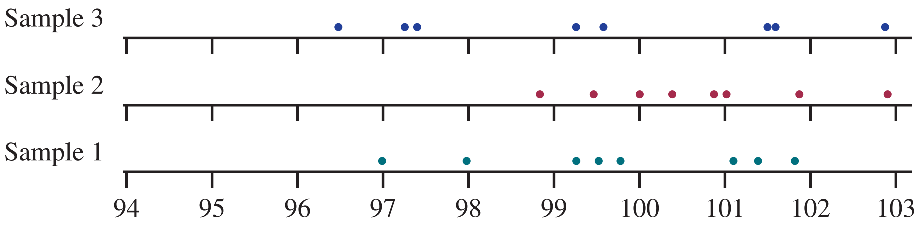
Activity 15.1 Exploring Single-Factor ANOVA

The purpose of this activity is to provide an opportunity to demonstrate your understanding of between-sample and within-sample variability and of the assumptions required for the ANOVA F test.

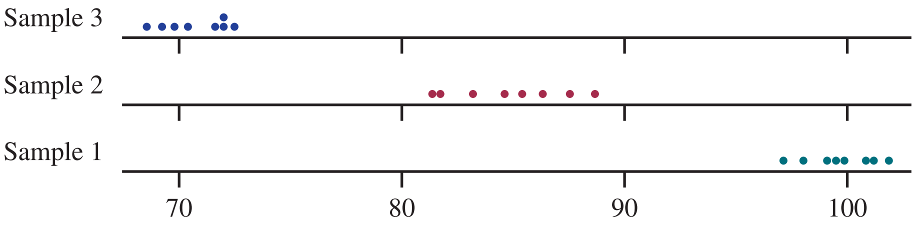
1. Each of the four graphs below show a dotplot of data from three separate random samples. For each of the four graphs, indicate whether you think that the basic assumptions for single-factor ANOVA are plausible. Write a sentence or two justifying your answer.



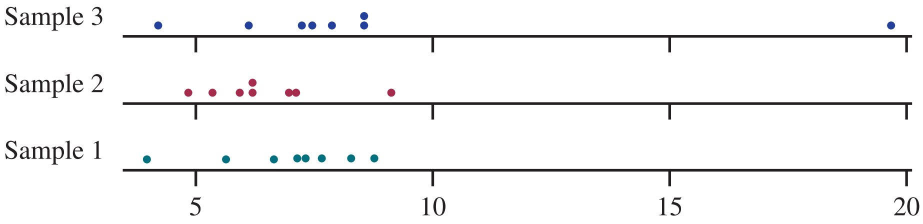
Graph 1



Graph 2



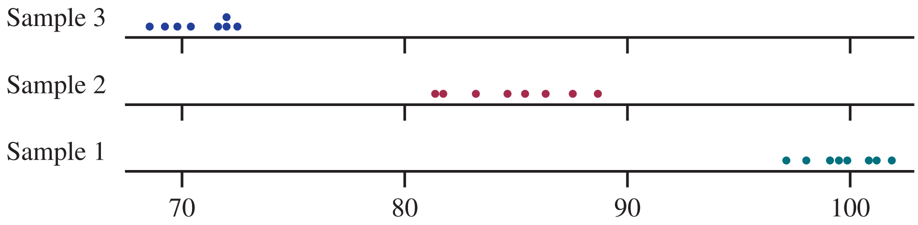
Graph 3



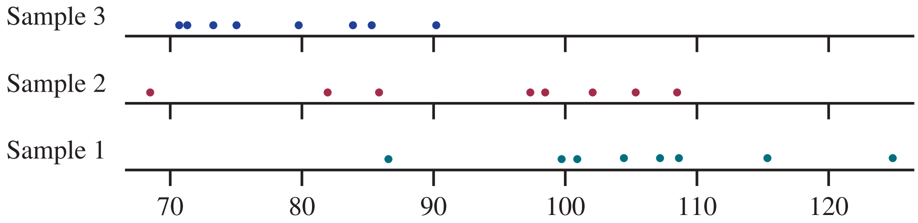
Graph 4

2. Each of the three graphs on the next page shows a dotplot of data from three separate random samples. For each of the three graphs, indicate whether you think that the three population means are probably not all the same, you think that the three population means might be the same, or you are unsure whether the population means could be the same. Write a sentence or two explaining your reasoning.
3. Sample data for each of the three graphs in Step 2 are shown in the table on the following page. For each of the three graphs, carry out a single-factor ANOVA. Are the results of the F tests consistent with your answers in Step 2? Explain.

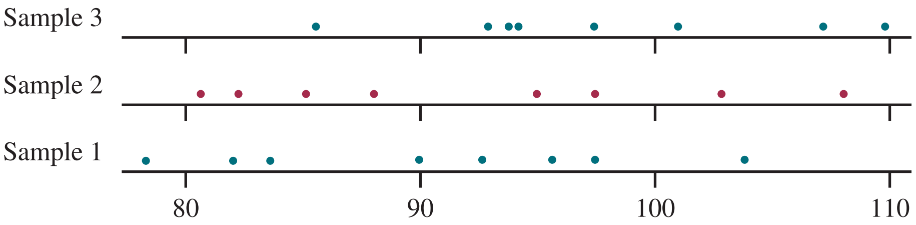
Copyright 2025 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s). Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.



Graph A



Graph B



Graph C

GRAPH A			GRAPH B			GRAPH C		
Sample 1	**Sample 2**	**Sample 3**	**Sample 1**	**Sample 2**	**Sample 3**	**Sample 1**	**Sample 2**	**Sample 3**
99.7	91.3	69.3	104.2	81.9	71.7	82.3	82.4	94.2
98.0	82.0	72.1	107.0	105.4	79.7	97.4	87.5	109.8
101.4	83.6	71.7	88.6	98.4	70.9	83.7	97.3	94.9
99.2	84.8	69.9	99.6	108.4	76.6	103.6	102.6	85.8
101.0	86.5	68.8	124.3	102.1	85.3	78.6	94.8	97.4
101.8	91.5	70.7	100.7	68.9	90.3	90.1	81.3	101.0
99.5	81.8	72.7	108.3	85.8	84.2	92.8	85.2	93.0
97.0	85.5	72.2	116.5	97.5	74.6	95.5	107.8	107.1

Summary Key Concepts and Formulas

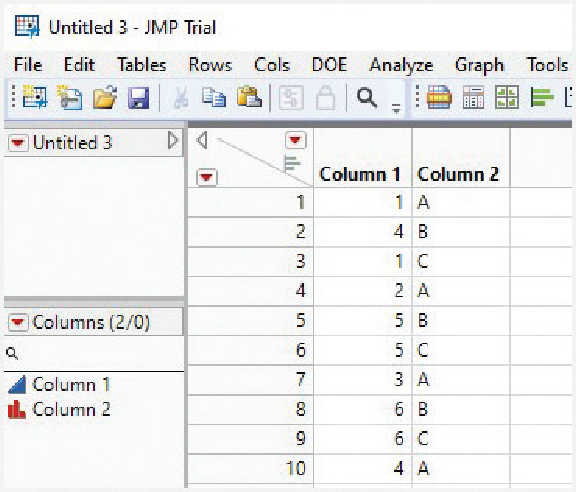
Term or Formula	Comment
Single-factor analysis of variance (ANOVA)	A test procedure for determining whether there are significant differences among k population or treatment means. The hypotheses tested are H_0: $\mu_1 = \mu_2 = \cdots = \mu_k$ versus H_a: at least two μ's differ.
Treatment sum of squares: $\mathbf{SSTr} = n_1(\bar{x}_1 - \bar{\bar{x}})^2 + \cdots + n_k(\bar{x}_k - \bar{\bar{x}})^2$	A measure of how different the k sample means $\bar{x}_1$, $\bar{x}_2, \ldots, \bar{x}_k$ are from one another; associated $\text{df}_1 = k - 1$.
Error sum of squares: $\mathbf{SSE} = (n_1 - 1)s_1^2 + \cdots + (n_k - 1)s_k^2$	A measure of the amount of variability within the individual samples; associated $\text{df}_2 = N - k$, where $N = n_1 + \cdots + n_k$.
Mean square	A sum of squares divided by its df. For single-factor ANOVA, $\text{MSTr} = \text{SSTr}/(k - 1)$ and $\text{MSE} = \text{SSE}/(N - k)$.
$F = \dfrac{\mathbf{MSTr}}{\mathbf{MSE}}$	The test statistic for testing H_0: $\mu_1 = \mu_2 = \cdots = \mu_k$ in a single-factor ANOVA. When H_0 is true, F has an F distribution with numerator $\text{df}_1 = k - 1$ and denominator $\text{df}_2 = N - k$.
$\mathbf{SSTo = SSTr + SSE}$	The fundamental identity in single-factor ANOVA, where SSTo = total sum of squares = $\Sigma(x - \bar{\bar{x}})^2$.
Tukey-Kramer multiple comparison procedure	A procedure for identifying significant differences among the μ's once the hypothesis H_0: $\mu_1 = \mu_2 = \cdots = \mu_k$ has been rejected by the ANOVA F test.

Technology Notes

ANOVA

JMP

1. Input the raw data into the first column
2. Input the group information into the second column



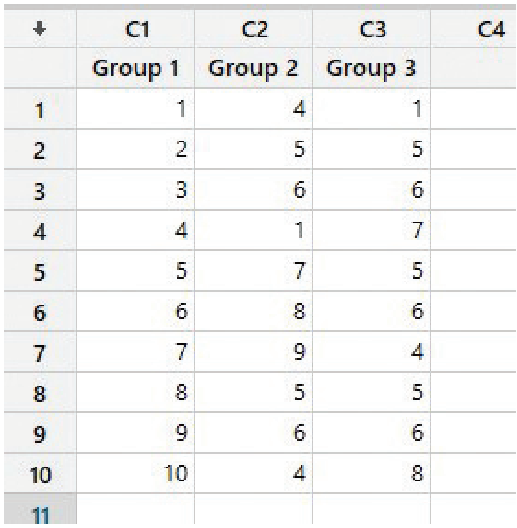
	Column 1	Column 2
1	1	A
2	4	B
3	1	C
4	2	A
5	5	B
6	5	C
7	3	A
8	6	B
9	6	C
10	4	A

3. Click **Analyze** then select **Fit Y by X**
4. Click and drag the first column name from the box under **Select Columns** to the box next to **Y, Response**
5. Click and drag the second column name from the box under **Select Columns** to the box next to **X, Factor**
6. Click **OK**
7. Click the red arrow next to **Oneway Analysis of…** and select **Means/ANOVA**

Minitab

Data stored in separate columns

1. Input each group's data in a separate column

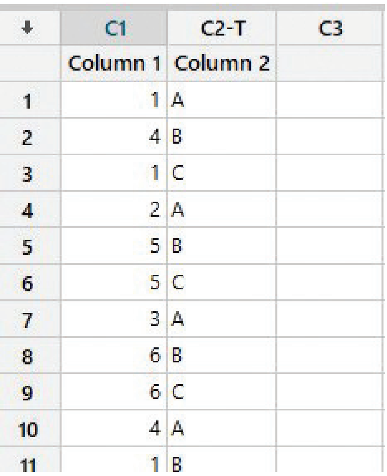


↓	C1	C2	C3	C4
	Group 1	Group 2	Group 3	
1	1	4	1	
2	2	5	5	
3	3	6	6	
4	4	1	7	
5	5	7	5	
6	6	8	6	
7	7	9	4	
8	8	5	5	
9	9	6	6	
10	10	4	8	
11				

2. Click **Stat** then **ANOVA** then **One-Way…**
3. Select **Response data are in a separate column for each factor level** from the dropdown menu
4. Click in the box under **Responses:**
5. Double-click the column name containing each group's data
6. Click **OK**

Data stored in one column

1. Input the data into one column
2. Input the group information into a second column

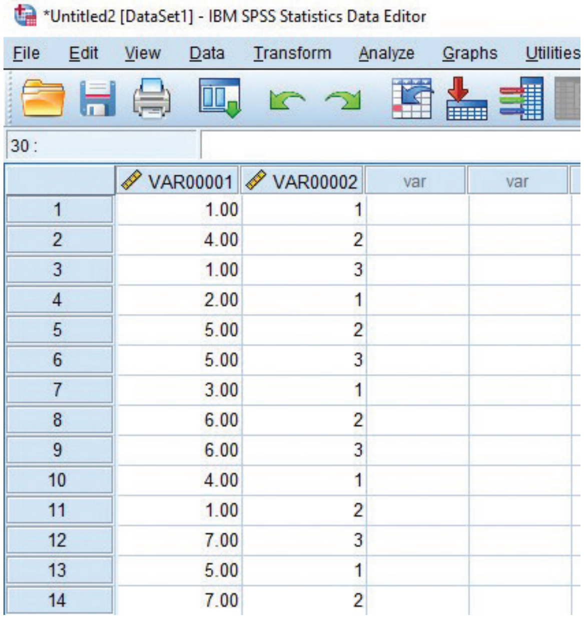


↓	C1	C2-T	C3
	Column 1	Column 2	
1	1	A	
2	4	B	
3	1	C	
4	2	A	
5	5	B	
6	5	C	
7	3	A	
8	6	B	
9	6	C	
10	4	A	
11	1	B	

3. Click **Stat** then **ANOVA** then **One-Way…**
4. Select **Response data are in one column for all factor levels** from the dropdown menu
5. Click in the box next to **Response:** and double-click the column name containing the raw data values
6. Click in the box next to **Factor:** and double-click the column name containing the group information
7. Click **OK**

SPSS

1. Input the raw data for all groups into one column
2. Input the group information into a second column (use group numbers)


*Untitled2 [DataSet1] - IBM SPSS Statistics Data Editor
File Edit View Data Transform Analyze Graphs Utilities
30 :

	VAR00001	VAR00002	var	var
1	1.00	1		
2	4.00	2		
3	1.00	3		
4	2.00	1		
5	5.00	2		
6	5.00	3		
7	3.00	1		
8	6.00	2		
9	6.00	3		
10	4.00	1		
11	1.00	2		
12	7.00	3		
13	5.00	1		
14	7.00	2		

Copyright 2025 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s). Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

3. Click **Analyze** then click **Compare Means** then click **One-Way ANOVA…**
4. Click the name of the column containing the raw data and click the arrow to move it to the box under **Dependent List:**
5. Click the name of the column containing the group data and click the arrow to move it to the box under **Factor:**
6. Click **OK**

Excel

1. Input the raw data for each group into a separate column
2. Click the **Data** ribbon
3. Click **Data Analysis** in the **Analysis** group

Note: If you do not see **Data Analysis** listed on the Ribbon, see the Technology Notes for Chapter 2 for instructions on installing this add-on.

4. Select **Anova: Single Factor** and click **OK**
5. Click on the box next to **Input Range** and select ALL columns of data (if you typed and selected column titles, click the box next to **Labels in First Row**)
6. Click in the box next to **Alpha** and type the significance level
7. Click **OK**

Note: The test statistic and p-value can be found in the first row of the ANOVA table under *F* and *P-value*, respectively.

TI-83/84

1. Enter the data for each group into a separate list starting with **L1** (In order to access lists press the **STAT** key, highlight the option called **Edit…** then press **ENTER**)
2. Press **STAT**
3. Highlight **TESTS**
4. Highlight **ANOVA** and press **ENTER**
5. Press **2nd** then **1**
6. Press**,**
7. Press **2nd** then **2**
8. Press**,**
9. Continue to input lists where data is stored separated by commas until you input the final list
10. When you are finished entering all lists, press **)**
11. Press **ENTER**

TI-Nspire

Summarized Data

1. Enter the summary statistics for the first group into a list in the following order: *n* in the first row, (insert the symbol for the sample mean x-bar) in the second row, and *s* in the third row (In order to access data lists select the spreadsheet option and press **enter**.)

Note: Be sure to title the lists by selecting the top row of the column and typing a title.

2. Continue to enter summary information for each group in this manner with each group in a separate list
3. When you are finished entering data for each group, press **menu** then **4:Statistics** then **4:Stat Tests** then **C:ANOVA…** then press **enter**
4. For **Data Input Method** choose **Stats** from the drop-down menu
5. For **Number of Groups** enter the number of groups, *k*
6. In the box next to **Group 1 Stats** select the list containing group one's summary statistics
7. In the box next to **Group 2 Stats** select the list containing group one's summary statistics
8. Continue entering summary statistics in this manner for all groups
9. Press **OK**

Raw data

1. Enter each group's data into separate data lists (In order to access data lists, select the spreadsheet option and press **enter**)

Note: Be sure to title the lists by selecting the top row of the column and typing a title.

2. Press the **menu** key and select **4:Statistics** then **4:Stat Tests** then **C:ANOVA…** and press **enter**
3. For **Data Input Method** choose **Data** from the drop-down menu
4. For **Number of Groups** input the number of groups, *k*
5. Press **OK**
6. For **List 1** select the list title that contains group one's data from the drop-down menu
7. For **List 2** select the list title that contains group two's data from the drop-down menu
8. Continue to select the appropriate lists for all groups
9. When you are finished inputting lists press **OK**

Appendix

Statistical Tables

Table 1 Random Numbers 822

Table 2 Standard Normal Probabilities (Cumulative z Curve Areas) 824

Table 3 t Critical Values 826

Table 4 Tail Areas for t Curves 827

Table 5 Curves of $\beta = P$(Type II Error) for t Tests 830

Table 6 Values That Capture Specified Upper-Tail F Curve Areas 831

Table 7 Critical Values of q for the Studentized Range Distribution 835

Table 8 Upper-Tail Areas for Chi-Square Distributions 836

Table 9 Binomial Probabilities 838

Appendix: Statistical Tables

Table 1 Random Numbers

Row																				
1	4	5	1	8	5	0	3	3	7	1	2	8	4	5	1	1	0	9	5	7
2	4	2	5	5	8	0	4	5	7	0	7	0	3	6	6	1	3	1	3	1
3	8	9	9	3	4	3	5	0	6	3	9	1	1	8	2	6	9	2	0	9
4	8	9	0	7	2	9	9	0	4	7	6	7	4	7	1	3	4	3	5	3
5	5	7	3	1	0	3	7	4	7	8	5	2	0	1	3	7	7	6	3	6
6	0	9	3	8	7	6	7	9	9	5	6	2	5	6	5	8	4	2	6	4
7	4	1	0	1	0	2	2	0	4	7	5	1	1	9	4	7	9	7	5	1
8	6	4	7	3	6	3	4	5	1	2	3	1	1	8	0	0	4	8	2	0
9	8	0	2	8	7	9	3	8	4	0	4	2	0	8	9	1	2	3	3	2
10	9	4	6	0	6	9	7	8	8	2	5	2	9	6	0	1	4	6	0	5
11	6	6	9	5	7	4	4	6	3	2	0	6	0	8	9	1	3	6	1	8
12	0	7	1	7	7	7	2	9	7	8	7	5	8	8	6	9	8	4	1	0
13	6	1	3	0	9	7	3	3	6	6	0	4	1	8	3	2	6	7	6	8
14	2	2	3	6	2	1	3	0	2	2	6	6	9	7	0	2	1	2	5	8
15	0	7	1	7	4	2	0	0	0	1	3	1	2	0	4	7	8	4	1	0
16	6	6	5	1	6	1	8	1	5	5	2	6	2	0	1	1	5	2	3	6
17	9	9	6	2	5	3	5	9	8	3	7	5	0	1	3	9	3	8	0	8
18	9	9	9	6	1	2	9	3	4	6	5	6	4	6	5	8	2	7	4	0
19	2	5	6	3	1	9	8	1	1	0	3	5	6	7	9	1	4	5	2	0
20	5	1	1	9	8	1	2	1	1	6	9	8	1	8	1	9	9	1	2	0
21	1	9	8	0	7	4	6	8	4	0	3	0	8	1	1	0	6	2	3	2
22	9	7	0	9	6	3	8	9	9	7	0	6	5	4	3	6	5	0	3	2
23	1	7	6	4	8	2	0	3	9	6	3	6	2	1	0	7	7	3	1	7
24	6	2	5	8	2	0	7	8	6	4	6	6	8	9	2	0	6	9	0	4
25	1	5	7	1	1	1	9	5	1	4	5	2	8	3	4	3	0	7	3	5
26	1	4	6	6	5	6	0	1	9	4	0	5	2	7	6	4	3	6	8	8
27	1	8	5	0	2	1	6	8	0	7	7	2	6	2	6	7	5	4	8	7
28	7	8	7	4	6	5	4	3	7	9	3	9	2	7	9	5	4	2	3	1
29	1	6	3	2	8	3	7	3	0	7	2	4	8	0	9	9	9	4	7	0
30	2	8	9	0	8	1	6	8	1	7	3	1	3	0	9	7	2	5	7	9
31	0	7	8	8	6	5	7	5	5	4	0	0	3	4	1	2	7	3	7	9
32	8	4	0	1	4	5	1	9	1	1	2	1	5	3	2	8	5	5	7	5
33	7	3	5	9	7	0	4	9	1	2	1	3	2	5	1	9	3	3	8	3
34	4	7	2	6	7	6	9	9	2	7	8	7	5	5	5	2	4	4	3	4
35	9	3	3	7	0	7	0	5	7	5	6	9	5	4	3	1	4	6	6	8
36	0	2	4	9	7	8	1	6	3	8	7	8	0	5	6	7	2	7	5	0
37	7	1	0	1	8	4	7	1	2	9	3	8	0	0	8	7	9	2	8	6
38	9	7	9	4	4	5	3	1	9	3	4	5	0	6	3	5	9	6	9	8
39	0	4	2	5	0	0	9	9	6	4	0	6	9	0	3	8	3	5	7	2
40	0	7	1	2	3	6	1	7	9	3	9	5	4	6	8	4	8	8	0	6
41	3	5	6	6	2	4	4	5	6	3	7	8	7	6	5	2	0	4	3	2
42	6	6	8	5	5	2	9	7	9	3	3	1	6	9	5	9	7	1	1	2
43	9	5	0	4	3	1	1	7	3	9	2	7	7	4	7	0	3	1	2	8
44	5	1	7	8	9	4	7	2	9	2	8	9	9	8	0	6	3	7	2	1
45	1	6	3	9	4	1	3	2	1	1	8	5	6	3	4	1	9	3	1	7
46	4	4	8	6	4	0	3	8	3	8	3	5	9	5	9	4	8	3	9	4
47	7	7	6	6	4	5	4	4	8	4	4	0	3	9	8	5	2	0	2	3
48	2	5	6	6	3	7	0	6	5	6	9	0	1	9	5	2	6	9	1	2

(continued)

Copyright 2025 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s). Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

Table 1 Random Numbers (Continued)

Row																				
49	9	4	0	4	7	5	3	2	8	7	2	7	4	9	3	9	6	5	5	6
50	7	3	1	5	6	6	5	0	3	5	3	7	2	8	6	2	4	1	8	7
51	7	5	8	2	8	8	8	7	6	4	1	1	0	2	3	1	9	3	6	0
52	3	3	6	0	9	1	1	0	3	2	7	8	2	0	5	3	4	8	9	8
53	0	2	9	6	9	8	9	3	8	1	5	3	9	9	7	0	7	7	1	6
54	8	5	9	6	2	9	6	8	2	1	2	4	7	0	6	8	3	4	6	1
55	5	4	7	6	1	0	0	1	0	4	6	1	4	1	5	0	9	6	5	5
56	5	0	3	6	4	1	9	8	4	4	1	2	0	2	5	1	8	1	2	1
57	0	2	6	3	7	5	1	1	6	6	0	5	8	1	2	3	3	6	1	3
58	3	8	1	6	3	8	1	4	5	2	9	4	2	5	7	3	2	3	1	8
59	9	1	5	6	0	6	5	6	6	3	6	2	3	0	0	0	1	8	5	9
60	5	3	5	6	3	9	5	4	7	3	6	6	7	5	0	1	5	6	7	3
61	9	6	6	4	5	7	7	6	1	5	4	4	8	0	6	5	7	6	3	0
62	6	3	0	6	7	9	5	5	4	6	2	2	8	4	4	0	0	9	9	8
63	8	5	8	3	5	2	0	6	6	0	0	6	0	6	3	0	1	7	0	5
64	3	8	2	4	9	0	9	2	6	2	9	5	1	9	1	9	0	8	3	3
65	1	4	4	1	1	7	4	6	3	6	5	6	5	5	7	7	0	3	5	8
66	5	9	9	5	3	7	2	5	1	7	1	1	0	7	1	0	9	2	8	8
67	8	7	1	7	5	2	5	6	8	7	9	9	1	3	9	6	4	9	3	0
68	6	7	2	3	1	4	9	2	1	7	0	8	6	7	8	9	9	4	7	4
69	2	3	2	8	7	0	9	7	1	1	1	2	8	2	9	1	0	6	7	7
70	2	9	5	7	8	4	7	9	0	3	6	9	2	0	6	0	6	2	6	8
71	4	8	9	8	3	2	7	6	9	1	9	8	6	9	5	2	4	9	9	9
72	1	5	6	5	7	7	5	4	3	4	3	8	1	8	9	9	4	4	1	1
73	1	8	1	1	7	2	8	5	5	8	9	9	9	6	2	0	1	6	6	7
74	5	7	7	0	9	5	5	6	8	6	8	2	2	6	0	5	5	1	8	7
75	1	8	6	0	5	4	8	3	4	5	3	5	8	7	7	7	8	5	7	0
76	2	6	6	7	9	4	2	2	8	7	4	3	4	9	6	1	9	4	3	9
77	3	6	6	4	5	7	8	3	0	2	8	4	6	7	2	1	4	5	2	3
78	0	7	8	0	1	2	1	1	3	4	2	1	6	9	3	3	5	4	0	4
79	8	3	6	0	5	7	7	9	1	5	8	8	4	9	5	7	2	2	7	6
80	5	3	6	9	0	6	3	8	7	5	9	5	9	7	4	2	5	6	2	9
81	0	9	3	7	7	2	8	6	4	3	2	9	4	8	2	9	9	6	9	9
82	9	4	7	4	0	0	0	3	5	4	6	6	2	6	2	3	6	1	1	4
83	5	5	4	1	7	8	6	4	2	3	2	9	8	4	6	3	8	3	0	5
84	5	3	0	0	5	4	8	0	7	4	7	6	2	1	1	2	1	2	6	9
85	3	3	0	9	3	2	9	4	0	5	5	4	8	7	5	7	5	3	8	8
86	3	0	5	7	1	9	5	8	0	0	4	5	3	0	3	0	2	7	6	7
87	5	0	8	6	0	8	1	6	2	0	8	6	5	4	0	7	2	9	1	0
88	3	6	4	7	8	2	3	5	7	9	8	5	2	7	6	9	0	2	4	9
89	9	0	4	4	9	1	6	8	5	2	8	9	0	7	5	7	2	5	1	8
90	9	5	2	6	9	3	9	6	5	1	8	8	7	8	2	0	4	4	7	9
91	9	4	5	7	0	3	4	6	4	2	5	4	8	6	1	1	9	1	8	8
92	8	1	1	8	0	5	4	2	8	5	3	3	3	0	1	1	4	4	8	3
93	6	9	4	7	8	3	3	9	1	2	5	0	1	2	3	0	1	1	2	5
94	0	0	6	8	8	7	2	4	4	7	6	6	0	3	4	7	5	6	8	2
95	5	3	3	9	3	8	4	9	1	9	1	7	8	4	5	2	2	5	4	4
96	2	5	6	2	7	6	0	3	8	1	4	4	2	6	8	3	6	3	2	8
97	7	4	3	7	9	6	8	6	2	8	3	8	4	2	2	0	7	0	5	3
98	1	9	0	8	8	0	1	2	2	2	7	5	6	5	5	7	8	7	2	6
99	2	4	8	0	2	5	2	7	0	5	9	6	6	1	5	8	7	9	7	5
100	4	1	7	8	6	7	1	1	5	8	9	4	8	9	8	3	0	9	0	7

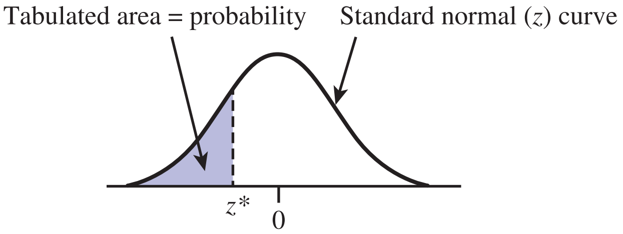


Table 2 **Standard Normal Probabilities (Cumulative z Curve Areas)**

z^*	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
−3.8	.0001	.0001	.0001	.0001	.0001	.0001	.0001	.0001	.0001	.0000
−3.7	.0001	.0001	.0001	.0001	.0001	.0001	.0001	.0001	.0001	.0001
−3.6	.0002	.0002	.0001	.0001	.0001	.0001	.0001	.0001	.0001	.0001
−3.5	.0002	.0002	.0002	.0002	.0002	.0002	.0002	.0002	.0002	.0002
−3.4	.0003	.0003	.0003	.0003	.0003	.0003	.0003	.0003	.0003	.0002
−3.3	.0005	.0005	.0005	.0004	.0004	.0004	.0004	.0004	.0004	.0003
−3.2	.0007	.0007	.0006	.0006	.0006	.0006	.0006	.0005	.0005	.0005
−3.1	.0010	.0009	.0009	.0009	.0008	.0008	.0008	.0008	.0007	.0007
−3.0	.0013	.0013	.0013	.0012	.0012	.0011	.0011	.0011	.0010	.0010
−2.9	.0019	.0018	.0018	.0017	.0016	.0016	.0015	.0015	.0014	.0014
−2.8	.0026	.0025	.0024	.0023	.0023	.0022	.0021	.0021	.0020	.0019
−2.7	.0035	.0034	.0033	.0032	.0031	.0030	.0029	.0028	.0027	.0026
−2.6	.0047	.0045	.0044	.0043	.0041	.0040	.0039	.0038	.0037	.0036
−2.5	.0062	.0060	.0059	.0057	.0055	.0054	.0052	.0051	.0049	.0048
−2.4	.0082	.0080	.0078	.0075	.0073	.0071	.0069	.0068	.0066	.0064
−2.3	.0107	.0104	.0102	.0099	.0096	.0094	.0091	.0089	.0087	.0084
−2.2	.0139	.0136	.0132	.0129	.0125	.0122	.0119	.0116	.0113	.0110
−2.1	.0179	.0174	.0170	.0166	.0162	.0158	.0154	.0150	.0146	.0143
−2.0	.0228	.0222	.0217	.0212	.0207	.0202	.0197	.0192	.0188	.0183
−1.9	.0287	.0281	.0274	.0268	.0262	.0256	.0250	.0244	.0239	.0233
−1.8	.0359	.0351	.0344	.0336	.0329	.0322	.0314	.0307	.0301	.0294
−1.7	.0446	.0436	.0427	.0418	.0409	.0401	.0392	.0384	.0375	.0367
−1.6	.0548	.0537	.0526	.0516	.0505	.0495	.0485	.0475	.0465	.0455
−1.5	.0668	.0655	.0643	.0630	.0618	.0606	.0594	.0582	.0571	.0559
−1.4	.0808	.0793	.0778	.0764	.0749	.0735	.0721	.0708	.0694	.0681
−1.3	.0968	.0951	.0934	.0918	.0901	.0885	.0869	.0853	.0838	.0823
−1.2	.1151	.1131	.1112	.1093	.1075	.1056	.1038	.1020	.1003	.0985
−1.1	.1357	.1335	.1314	.1292	.1271	.1251	.1230	.1210	.1190	.1170
−1.0	.1587	.1562	.1539	.1515	.1492	.1469	.1446	.1423	.1401	.1379
−0.9	.1841	.1814	.1788	.1762	.1736	.1711	.1685	.1660	.1635	.1611
−0.8	.2119	.2090	.2061	.2033	.2005	.1977	.1949	.1922	.1894	.1867
−0.7	.2420	.2389	.2358	.2327	.2296	.2266	.2236	.2206	.2177	.2148
−0.6	.2743	.2709	.2676	.2643	.2611	.2578	.2546	.2514	.2483	.2451
−0.5	.3085	.3050	.3015	.2981	.2946	.2912	.2877	.2843	.2810	.2776
−0.4	.3446	.3409	.3372	.3336	.3300	.3264	.3228	.3192	.3156	.3121
−0.3	.3821	.3783	.3745	.3707	.3669	.3632	.3594	.3557	.3520	.3483
−0.2	.4207	.4168	.4129	.4090	.4052	.4013	.3974	.3936	.3897	.3859
−0.1	.4602	.4562	.4522	.4483	.4443	.4404	.4364	.4325	.4286	.4247
−0.0	.5000	.4960	.4920	.4880	.4840	.4801	.4761	.4721	.4681	.4641

(continued)

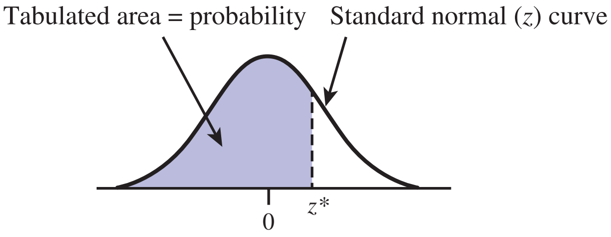


Table 2 **Standard Normal Probabilities (Cumulative *z* Curve Areas) (Continued)**

*z**	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
0.0	.5000	.5040	.5080	.5120	.5160	.5199	.5239	.5279	.5319	.5359
0.1	.5398	.5438	.5478	.5517	.5557	.5596	.5636	.5675	.5714	.5753
0.2	.5793	.5832	.5871	.5910	.5948	.5987	.6026	.6064	.6103	.6141
0.3	.6179	.6217	.6255	.6293	.6331	.6368	.6406	.6443	.6480	.6517
0.4	.6554	.6591	.6628	.6664	.6700	.6736	.6772	.6808	.6844	.6879
0.5	.6915	.6950	.6985	.7019	.7054	.7088	.7123	.7157	.7190	.7224
0.6	.7257	.7291	.7324	.7357	.7389	.7422	.7454	.7486	.7517	.7549
0.7	.7580	.7611	.7642	.7673	.7704	.7734	.7764	.7794	.7823	.7852
0.8	.7881	.7910	.7939	.7967	.7995	.8023	.8051	.8078	.8106	.8133
0.9	.8159	.8186	.8212	.8238	.8264	.8289	.8315	.8340	.8365	.8389
1.0	.8413	.8438	.8461	.8485	.8508	.8531	.8554	.8577	.8599	.8621
1.1	.8643	.8665	.8686	.8708	.8729	.8749	.8770	.8790	.8810	.8830
1.2	.8849	.8869	.8888	.8907	.8925	.8944	.8962	.8980	.8997	.9015
1.3	.9032	.9049	.9066	.9082	.9099	.9115	.9131	.9147	.9162	.9177
1.4	.9192	.9207	.9222	.9236	.9251	.9265	.9279	.9292	.9306	.9319
1.5	.9332	.9345	.9357	.9370	.9382	.9394	.9406	.9418	.9429	.9441
1.6	.9452	.9463	.9474	.9484	.9495	.9505	.9515	.9525	.9535	.9545
1.7	.9554	.9564	.9573	.9582	.9591	.9599	.9608	.9616	.9625	.9633
1.8	.9641	.9649	.9656	.9664	.9671	.9678	.9686	.9693	.9699	.9706
1.9	.9713	.9719	.9726	.9732	.9738	.9744	.9750	.9756	.9761	.9767
2.0	.9772	.9778	.9783	.9788	.9793	.9798	.9803	.9808	.9812	.9817
2.1	.9821	.9826	.9830	.9834	.9838	.9842	.9846	.9850	.9854	.9857
2.2	.9861	.9864	.9868	.9871	.9875	.9878	.9881	.9884	.9887	.9890
2.3	.9893	.9896	.9898	.9901	.9904	.9906	.9909	.9911	.9913	.9916
2.4	.9918	.9920	.9922	.9925	.9927	.9929	.9931	.9932	.9934	.9936
2.5	.9938	.9940	.9941	.9943	.9945	.9946	.9948	.9949	.9951	.9952
2.6	.9953	.9955	.9956	.9957	.9959	.9960	.9961	.9962	.9963	.9964
2.7	.9965	.9966	.9967	.9968	.9969	.9970	.9971	.9972	.9973	.9974
2.8	.9974	.9975	.9976	.9977	.9977	.9978	.9979	.9979	.9980	.9981
2.9	.9981	.9982	.9982	.9983	.9984	.9984	.9985	.9985	.9986	.9986
3.0	.9987	.9987	.9987	.9988	.9988	.9989	.9989	.9989	.9990	.9990
3.1	.9990	.9991	.9991	.9991	.9992	.9992	.9992	.9992	.9993	.9993
3.2	.9993	.9993	.9994	.9994	.9994	.9994	.9994	.9995	.9995	.9995
3.3	.9995	.9995	.9995	.9996	.9996	.9996	.9996	.9996	.9996	.9997
3.4	.9997	.9997	.9997	.9997	.9997	.9997	.9997	.9997	.9997	.9998
3.5	.9998	.9998	.9998	.9998	.9998	.9998	.9998	.9998	.9998	.9998
3.6	.9998	.9998	.9999	.9999	.9999	.9999	.9999	.9999	.9999	.9999
3.7	.9999	.9999	.9999	.9999	.9999	.9999	.9999	.9999	.9999	.9999
3.8	.9999	.9999	.9999	.9999	.9999	.9999	.9999	.9999	.9999	1.0000

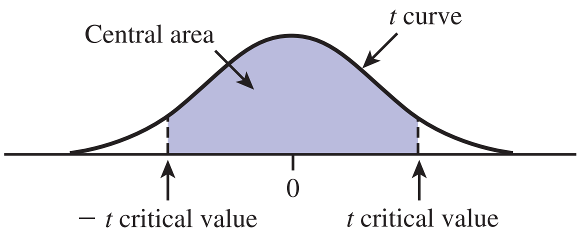


Table 3 *t* Critical Values

Central area captured:		**.80**	**.90**	**.95**	**.98**	**.99**	**.998**	**.999**
Confidence level:		**80%**	**90%**	**95%**	**98%**	**99%**	**99.8%**	**99.9%**
Degrees of freedom	**1**	3.08	6.31	12.71	31.82	63.66	318.31	636.62
	2	1.89	2.92	4.30	6.97	9.93	23.33	31.60
	3	1.64	2.35	3.18	4.54	5.84	10.21	12.92
	4	1.53	2.13	2.78	3.75	4.60	7.17	8.61
	5	1.48	2.02	2.57	3.37	4.03	5.89	6.86
	6	1.44	1.94	2.45	3.14	3.71	5.21	5.96
	7	1.42	1.90	2.37	3.00	3.50	4.79	5.41
	8	1.40	1.86	2.31	2.90	3.36	4.50	5.04
	9	1.38	1.83	2.26	2.82	3.25	4.30	4.78
	10	1.37	1.81	2.23	2.76	3.17	4.14	4.59
	11	1.36	1.80	2.20	2.72	3.11	4.03	4.44
	12	1.36	1.78	2.18	2.68	3.06	3.93	4.32
	13	1.35	1.77	2.16	2.65	3.01	3.85	4.22
	14	1.35	1.76	2.15	2.62	2.98	3.79	4.14
	15	1.34	1.75	2.13	2.60	2.95	3.73	4.07
	16	1.34	1.75	2.12	2.58	2.92	3.69	4.02
	17	1.33	1.74	2.11	2.57	2.90	3.65	3.97
	18	1.33	1.73	2.10	2.55	2.88	3.61	3.92
	19	1.33	1.73	2.09	2.54	2.86	3.58	3.88
	20	1.33	1.73	2.09	2.53	2.85	3.55	3.85
	21	1.32	1.72	2.08	2.52	2.83	3.53	3.82
	22	1.32	1.72	2.07	2.51	2.82	3.51	3.79
	23	1.32	1.71	2.07	2.50	2.81	3.49	3.77
	24	1.32	1.71	2.06	2.49	2.80	3.47	3.75
	25	1.32	1.71	2.06	2.49	2.79	3.45	3.73
	26	1.32	1.71	2.06	2.48	2.78	3.44	3.71
	27	1.31	1.70	2.05	2.47	2.77	3.42	3.69
	28	1.31	1.70	2.05	2.47	2.76	3.41	3.67
	29	1.31	1.70	2.05	2.46	2.76	3.40	3.66
	30	1.31	1.70	2.04	2.46	2.75	3.39	3.65
	40	1.30	1.68	2.02	2.42	2.70	3.31	3.55
	60	1.30	1.67	2.00	2.39	2.66	3.23	3.46
	120	1.29	1.66	1.98	2.36	2.62	3.16	3.37
***z* critical values**	**∞**	1.28	1.645	1.96	2.33	2.58	3.09	3.29

Copyright 2025 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s). Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

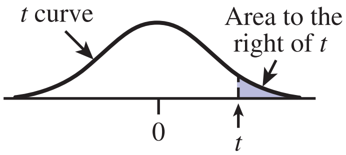


Table 4 Tail Areas for *t* Curves

t \ df	1	2	3	4	5	6	7	8	9	10	11	12
0.0	.500	.500	.500	.500	.500	.500	.500	.500	.500	.500	.500	.500
0.1	.468	.465	.463	.463	.462	.462	.462	.461	.461	.461	.461	.461
0.2	.437	.430	.427	.426	.425	.424	.424	.423	.423	.423	.423	.422
0.3	.407	.396	.392	.390	.388	.387	.386	.386	.386	.385	.385	.385
0.4	.379	.364	.358	.355	.353	.352	.351	.350	.349	.349	.348	.348
0.5	.352	.333	.326	.322	.319	.317	.316	.315	.315	.314	.313	.313
0.6	.328	.305	.295	.290	.287	.285	.284	.283	.282	.281	.280	.280
0.7	.306	.278	.267	.261	.258	.255	.253	.252	.251	.250	.249	.249
0.8	.285	.254	.241	.234	.230	.227	.225	.223	.222	.221	.220	.220
0.9	.267	.232	.217	.210	.205	.201	.199	.197	.196	.195	.194	.193
1.0	.250	.211	.196	.187	.182	.178	.175	.173	.172	.170	.169	.169
1.1	.235	.193	.176	.167	.162	.157	.154	.152	.150	.149	.147	.146
1.2	.221	.177	.158	.148	.142	.138	.135	.132	.130	.129	.128	.127
1.3	.209	.162	.142	.132	.125	.121	.117	.115	.113	.111	.110	.109
1.4	.197	.148	.128	.117	.110	.106	.102	.100	.098	.096	.095	.093
1.5	.187	.136	.115	.104	.097	.092	.089	.086	.084	.082	.081	.080
1.6	.178	.125	.104	.092	.085	.080	.077	.074	.072	.070	.069	.068
1.7	.169	.116	.094	.082	.075	.070	.066	.064	.062	.060	.059	.057
1.8	.161	.107	.085	.073	.066	.061	.057	.055	.053	.051	.050	.049
1.9	.154	.099	.077	.065	.058	.053	.050	.047	.045	.043	.042	.041
2.0	.148	.092	.070	.058	.051	.046	.043	.040	.038	.037	.035	.034
2.1	.141	.085	.063	.052	.045	.040	.037	.034	.033	.031	.030	.029
2.2	.136	.079	.058	.046	.040	.035	.032	.029	.028	.026	.025	.024
2.3	.131	.074	.052	.041	.035	.031	.027	.025	.023	.022	.021	.020
2.4	.126	.069	.048	.037	.031	.027	.024	.022	.020	.019	.018	.017
2.5	.121	.065	.044	.033	.027	.023	.020	.018	.017	.016	.015	.014
2.6	.117	.061	.040	.030	.024	.020	.018	.016	.014	.013	.012	.012
2.7	.113	.057	.037	.027	.021	.018	.015	.014	.012	.011	.010	.010
2.8	.109	.054	.034	.024	.019	.016	.013	.012	.010	.009	.009	.008
2.9	.106	.051	.031	.022	.017	.014	.011	.010	.009	.008	.007	.007
3.0	.102	.048	.029	.020	.015	.012	.010	.009	.007	.007	.006	.006
3.1	.099	.045	.027	.018	.013	.011	.009	.007	.006	.006	.005	.005
3.2	.096	.043	.025	.016	.012	.009	.008	.006	.005	.005	.004	.004
3.3	.094	.040	.023	.015	.011	.008	.007	.005	.005	.004	.004	.003
3.4	.091	.038	.021	.014	.010	.007	.006	.005	.004	.003	.003	.003
3.5	.089	.036	.020	.012	.009	.006	.005	.004	.003	.003	.002	.002
3.6	.086	.035	.018	.011	.008	.006	.004	.004	.003	.002	.002	.002
3.7	.084	.033	.017	.010	.007	.005	.004	.003	.002	.002	.002	.002
3.8	.082	.031	.016	.010	.006	.004	.003	.003	.002	.002	.001	.001
3.9	.080	.030	.015	.009	.006	.004	.003	.002	.002	.001	.001	.001
4.0	.078	.029	.014	.008	.005	.004	.003	.002	.002	.001	.001	.001

(continued)

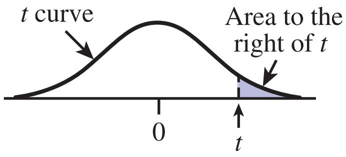


Table 4 Tail Areas for *t* Curves (Continued)

t \ df	13	14	15	16	17	18	19	20	21	22	23	24
0.0	.500	.500	.500	.500	.500	.500	.500	.500	.500	.500	.500	.500
0.1	.461	.461	.461	.461	.461	.461	.461	.461	.461	.461	.461	.461
0.2	.422	.422	.422	.422	.422	.422	.422	.422	.422	.422	.422	.422
0.3	.384	.384	.384	.384	.384	.384	.384	.384	.384	.383	.383	.383
0.4	.348	.347	.347	.347	.347	.347	.347	.347	.347	.347	.346	.346
0.5	.313	.312	.312	.312	.312	.312	.311	.311	.311	.311	.311	.311
0.6	.279	.279	.279	.278	.278	.278	.278	.278	.278	.277	.277	.277
0.7	.248	.247	.247	.247	.247	.246	.246	.246	.246	.246	.245	.245
0.8	.219	.218	.218	.218	.217	.217	.217	.217	.216	.216	.216	.216
0.9	.192	.191	.191	.191	.190	.190	.190	.189	.189	.189	.189	.189
1.0	.168	.167	.167	.166	.166	.165	.165	.165	.164	.164	.164	.164
1.1	.146	.144	.144	.144	.143	.143	.143	.142	.142	.142	.141	.141
1.2	.126	.124	.124	.124	.123	.123	.122	.122	.122	.121	.121	.121
1.3	.108	.107	.107	.106	.105	.105	.105	.104	.104	.104	.103	.103
1.4	.092	.091	.091	.090	.090	.089	.089	.089	.088	.088	.087	.087
1.5	.079	.077	.077	.077	.076	.075	.075	.075	.074	.074	.074	.073
1.6	.067	.065	.065	.065	.064	.064	.063	.063	.062	.062	.062	.061
1.7	.056	.055	.055	.054	.054	.053	.053	.052	.052	.052	.051	.051
1.8	.048	.046	.046	.045	.045	.044	.044	.043	.043	.043	.042	.042
1.9	.040	.038	.038	.038	.037	.037	.036	.036	.036	.035	.035	.035
2.0	.033	.032	.032	.031	.031	.030	.030	.030	.029	.029	.029	.028
2.1	.028	.027	.027	.026	.025	.025	.025	.024	.024	.024	.023	.023
2.2	.023	.022	.022	.021	.021	.021	.020	.020	.020	.019	.019	.019
2.3	.019	.018	.018	.018	.017	.017	.016	.016	.016	.016	.015	.015
2.4	.016	.015	.015	.014	.014	.014	.013	.013	.013	.013	.012	.012
2.5	.013	.012	.012	.012	.011	.011	.011	.011	.010	.010	.010	.010
2.6	.011	.010	.010	.010	.009	.009	.009	.009	.008	.008	.008	.008
2.7	.009	.008	.008	.008	.008	.007	.007	.007	.007	.007	.006	.006
2.8	.008	.007	.007	.006	.006	.006	.006	.006	.005	.005	.005	.005
2.9	.006	.005	.005	.005	.005	.005	.005	.004	.004	.004	.004	.004
3.0	.005	.004	.004	.004	.004	.004	.004	.004	.003	.003	.003	.003
3.1	.004	.004	.004	.003	.003	.003	.003	.003	.003	.003	.003	.002
3.2	.003	.003	.003	.003	.003	.002	.002	.002	.002	.002	.002	.002
3.3	.003	.002	.002	.002	.002	.002	.002	.002	.002	.002	.002	.001
3.4	.002	.002	.002	.002	.002	.002	.002	.001	.001	.001	.001	.001
3.5	.002	.002	.002	.001	.001	.001	.001	.001	.001	.001	.001	.001
3.6	.002	.001	.001	.001	.001	.001	.001	.001	.001	.001	.001	.001
3.7	.001	.001	.001	.001	.001	.001	.001	.001	.001	.001	.001	.001
3.8	.001	.001	.001	.001	.001	.001	.001	.001	.001	.000	.000	.000
3.9	.001	.001	.001	.001	.001	.001	.000	.000	.000	.000	.000	.000
4.0	.001	.001	.001	.001	.000	.000	.000	.000	.000	.000	.000	.000

(continued)

Copyright 2025 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s). Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

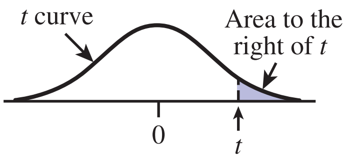
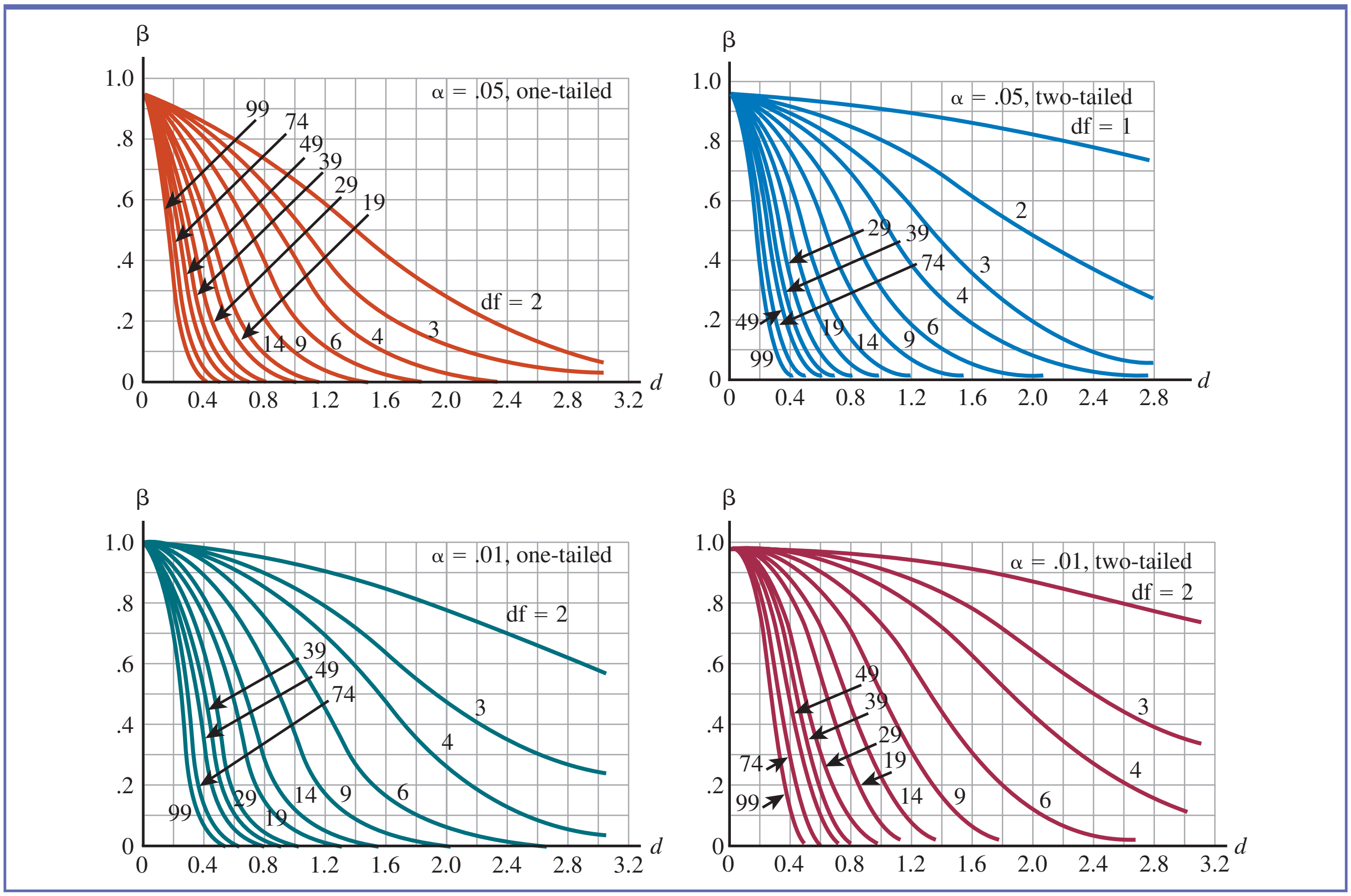


Table 4 Tail Areas for *t* Curves (Continued)

t \ df	25	26	27	28	29	30	35	40	60	120	∞(=z)
0.0	.500	.500	.500	.500	.500	.500	.500	.500	.500	.500	.500
0.1	.461	.461	.461	.461	.461	.461	.460	.460	.460	.460	.460
0.2	.422	.422	.421	.421	.421	.421	.421	.421	.421	.421	.421
0.3	.383	.383	.383	.383	.383	.383	.383	.383	.383	.382	.382
0.4	.346	.346	.346	.346	.346	.346	.346	.346	.345	.345	.345
0.5	.311	.311	.311	.310	.310	.310	.310	.310	.309	.309	.309
0.6	.277	.277	.277	.277	.277	.277	.276	.276	.275	.275	.274
0.7	.245	.245	.245	.245	.245	.245	.244	.244	.243	.243	.242
0.8	.216	.215	.215	.215	.215	.215	.215	.214	.213	.213	.212
0.9	.188	.188	.188	.188	.188	.188	.187	.187	.186	.185	.184
1.0	.163	.163	.163	.163	.163	.163	.162	.162	.161	.160	.159
1.1	.141	.141	.141	.140	.140	.140	.139	.139	.138	.137	.136
1.2	.121	.120	.120	.120	.120	.120	.119	.119	.117	.116	.115
1.3	.103	.103	.102	.102	.102	.102	.101	.101	.099	.098	.097
1.4	.087	.087	.086	.086	.086	.086	.085	.085	.083	.082	.081
1.5	.073	.073	.073	.072	.072	.072	.071	.071	.069	.068	.067
1.6	.061	.061	.061	.060	.060	.060	.059	.059	.057	.056	.055
1.7	.051	.051	.050	.050	.050	.050	.049	.048	.047	.046	.045
1.8	.042	.042	.042	.041	.041	.041	.040	.040	.038	.037	.036
1.9	.035	.034	.034	.034	.034	.034	.033	.032	.031	.030	.029
2.0	.028	.028	.028	.028	.027	.027	.027	.026	.025	.024	.023
2.1	.023	.023	.023	.022	.022	.022	.022	.021	.020	.019	.018
2.2	.019	.018	.018	.018	.018	.018	.017	.017	.016	.015	.014
2.3	.015	.015	.015	.015	.014	.014	.014	.013	.012	.012	.011
2.4	.012	.012	.012	.012	.012	.011	.011	.011	.010	.009	.008
2.5	.010	.010	.009	.009	.009	.009	.009	.008	.008	.007	.006
2.6	.008	.008	.007	.007	.007	.007	.007	.007	.006	.005	.005
2.7	.006	.006	.006	.006	.006	.006	.005	.005	.004	.004	.003
2.8	.005	.005	.005	.005	.005	.004	.004	.004	.003	.003	.003
2.9	.004	.004	.004	.004	.004	.003	.003	.003	.003	.002	.002
3.0	.003	.003	.003	.003	.003	.003	.002	.002	.002	.002	.001
3.1	.002	.002	.002	.002	.002	.002	.002	.002	.001	.001	.001
3.2	.002	.002	.002	.002	.002	.002	.001	.001	.001	.001	.001
3.3	.001	.001	.001	.001	.001	.001	.001	.001	.001	.001	.000
3.4	.001	.001	.001	.001	.001	.001	.001	.001	.001	.000	.000
3.5	.001	.001	.001	.001	.001	.001	.001	.001	.000	.000	.000
3.6	.001	.001	.001	.001	.001	.001	.000	.000	.000	.000	.000
3.7	.001	.001	.000	.000	.000	.000	.000	.000	.000	.000	.000
3.8	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000
3.9	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000
4.0	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000

Table 5 **Curves of $\beta = P$(Type II Error) for t Tests**



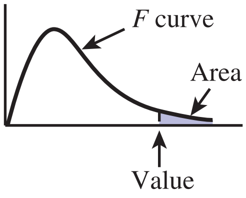


Table 6 Values That Capture Specified Upper-Tail *F* Curve Areas

		df_1									
df_2	Area	1	2	3	4	5	6	7	8	9	10
1	.10	39.86	49.50	53.59	55.83	57.24	58.20	58.91	59.44	59.86	60.19
	.05	161.40	199.50	215.70	224.60	230.20	234.00	236.80	238.90	240.50	241.90
	.01	4052.00	5000.00	5403.00	5625.00	5764.00	5859.00	5928.00	5981.00	6022.00	6056.00
2	.10	8.53	9.00	9.16	9.24	9.29	9.33	9.35	9.37	9.38	9.39
	.05	18.51	19.00	19.16	19.25	19.30	19.33	19.35	19.37	19.38	19.40
	.01	98.50	99.00	99.17	99.25	99.30	99.33	99.36	99.37	99.39	99.40
	.001	998.50	999.00	999.20	999.20	999.30	999.30	999.40	999.40	999.40	999.40
3	.10	5.54	5.46	5.39	5.34	5.31	5.28	5.27	5.25	5.24	5.23
	.05	10.13	9.55	9.28	9.12	9.01	8.94	8.89	8.85	8.81	8.79
	.01	34.12	30.82	29.46	28.71	28.24	27.91	27.67	27.49	27.35	27.23
	.001	167.00	148.50	141.10	137.10	134.60	132.80	131.60	130.60	129.90	129.20
4	.10	4.54	4.32	4.19	4.11	4.05	4.01	3.98	3.95	3.94	3.92
	.05	7.71	6.94	6.59	6.39	6.26	6.16	6.09	6.04	6.00	5.96
	.01	21.20	18.00	16.69	15.98	15.52	15.21	14.98	14.80	14.66	14.55
	.001	74.14	61.25	56.18	53.44	51.71	50.53	49.66	49.00	48.47	48.05
5	.10	4.06	3.78	3.62	3.52	3.45	3.40	3.37	3.34	3.32	3.30
	.05	6.61	5.79	5.41	5.19	5.05	4.95	4.88	4.82	4.77	4.74
	.01	16.26	13.27	12.06	11.39	10.97	10.67	10.46	10.29	10.16	10.05
	.001	47.18	37.12	33.20	31.09	29.75	28.83	28.16	27.65	27.24	26.92
6	.10	3.78	3.46	3.29	3.18	3.11	3.05	3.01	2.98	2.96	2.94
	.05	5.99	5.14	4.76	4.53	4.39	4.28	4.21	4.15	4.10	4.06
	.01	13.75	10.92	9.78	9.15	8.75	8.47	8.26	8.10	7.98	7.87
	.001	35.51	27.00	23.70	21.92	20.80	20.03	19.46	19.03	18.69	18.41
7	.10	3.59	3.26	3.07	2.96	2.88	2.83	2.78	2.75	2.72	2.70
	.05	5.59	4.74	4.35	4.12	3.97	3.87	3.79	3.73	3.68	3.64
	.01	12.25	9.55	8.45	7.85	7.46	7.19	6.99	6.84	6.72	6.62
	.001	29.25	21.69	18.77	17.20	16.21	15.52	15.02	14.63	14.33	14.08
8	.10	3.46	3.11	2.92	2.81	2.73	2.67	2.62	2.59	2.56	2.54
	.05	5.32	4.46	4.07	3.84	3.69	3.58	3.50	3.44	3.39	3.35
	.01	11.26	8.65	7.59	7.01	6.63	6.37	6.18	6.03	5.91	5.81
	.001	25.41	18.49	15.83	14.39	13.48	12.86	12.40	12.05	11.77	11.54
9	.10	3.36	3.01	2.81	2.69	2.61	2.55	2.51	2.47	2.44	2.42
	.05	5.12	4.26	3.86	3.63	3.48	3.37	3.29	3.23	3.18	3.14
	.01	10.56	8.02	6.99	6.42	6.06	5.80	5.61	5.47	5.35	5.26
	.001	22.86	16.39	13.90	12.56	11.71	11.13	10.70	10.37	10.11	9.89

(continued)

Table 6 Values That Capture Specified Upper-Tail *F* Curve Areas (Continued)

		df_1									
df_2	Area	1	2	3	4	5	6	7	8	9	10
10	.10	3.29	2.92	2.73	2.61	2.52	2.46	2.41	2.38	2.35	2.32
	.05	4.96	4.10	3.71	3.48	3.33	3.22	3.14	3.07	3.02	2.98
	.01	10.04	7.56	6.55	5.99	5.64	5.39	5.20	5.06	4.94	4.85
	.001	21.04	14.91	12.55	11.28	10.48	9.93	9.52	9.20	8.96	8.75
11	.10	3.23	2.86	2.66	2.54	2.45	2.39	2.34	2.30	2.27	2.25
	.05	4.84	3.98	3.59	3.36	3.20	3.09	3.01	2.95	2.90	2.85
	.01	9.65	7.21	6.22	5.67	5.32	5.07	4.89	4.74	4.63	4.54
	.001	19.69	13.81	11.56	10.35	9.58	9.05	8.66	8.35	8.12	7.92
12	.10	3.18	2.81	2.61	2.48	2.39	2.33	2.28	2.24	2.21	2.19
	.05	4.75	3.89	3.49	3.26	3.11	3.00	2.91	2.85	2.80	2.75
	.01	9.33	6.93	5.95	5.41	5.06	4.82	4.64	4.50	4.39	4.30
	.001	18.64	12.97	10.80	9.63	8.89	8.38	8.00	7.71	7.48	7.29
13	.10	3.14	2.76	2.56	2.43	2.35	2.28	2.23	2.20	2.16	2.14
	.05	4.67	3.81	3.41	3.18	3.03	2.92	2.83	2.77	2.71	2.67
	.01	9.07	6.70	5.74	5.21	4.86	4.62	4.44	4.30	4.19	4.10
	.001	17.82	12.31	10.21	9.07	8.35	7.86	7.49	7.21	6.98	6.80
14	.10	3.10	2.73	2.52	2.39	2.31	2.24	2.19	2.15	2.12	2.10
	.05	4.60	3.74	3.34	3.11	2.96	2.85	2.76	2.70	2.65	2.60
	.01	8.86	6.51	5.56	5.04	4.69	4.46	4.28	4.14	4.03	3.94
	.001	17.14	11.78	9.73	8.62	7.92	7.44	7.08	6.80	6.58	6.40
15	.10	3.07	2.70	2.49	2.36	2.27	2.21	2.16	2.12	2.09	2.06
	.05	4.54	3.68	3.29	3.06	2.90	2.79	2.71	2.64	2.59	2.54
	.01	8.68	6.36	5.42	4.89	4.56	4.32	4.14	4.00	3.89	3.80
	.001	16.59	11.34	9.34	8.25	7.57	7.09	6.74	6.47	6.26	6.08
16	.10	3.05	2.67	2.46	2.33	2.24	2.18	2.13	2.09	2.06	2.03
	.05	4.49	3.63	3.24	3.01	2.85	2.74	2.66	2.59	2.54	2.49
	.01	8.53	6.23	5.29	4.77	4.44	4.20	4.03	3.89	3.78	3.69
	.001	16.12	10.97	9.01	7.94	7.27	6.80	6.46	6.19	5.98	5.81
17	.10	3.03	2.64	2.44	2.31	2.22	2.15	2.10	2.06	2.03	2.00
	.05	4.45	3.59	3.20	2.96	2.81	2.70	2.61	2.55	2.49	2.45
	.01	8.40	6.11	5.18	4.67	4.34	4.10	3.93	3.79	3.68	3.59
	.001	15.72	10.66	8.73	7.68	7.02	6.56	6.22	5.96	5.75	5.58
18	.10	3.01	2.62	2.42	2.29	2.20	2.13	2.08	2.04	2.00	1.98
	.05	4.41	3.55	3.16	2.93	2.77	2.66	2.58	2.51	2.46	2.41
	.01	8.29	6.01	5.09	4.58	4.25	4.01	3.84	3.71	3.60	3.51
	.001	15.38	10.39	8.49	7.46	6.81	6.35	6.02	5.76	5.56	5.39
19	.10	2.99	2.61	2.40	2.27	2.18	2.11	2.06	2.02	1.98	1.96
	.05	4.38	3.52	3.13	2.90	2.74	2.63	2.54	2.48	2.42	2.38
	.01	8.18	5.93	5.01	4.50	4.17	3.94	3.77	3.63	3.52	3.43
	.001	15.08	10.16	8.28	7.27	6.62	6.18	5.85	5.59	5.39	5.22

(continued)

Copyright 2025 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s). Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

Table 6 **Values That Capture Specified Upper-Tail *F* Curve Areas (Continued)**

		df$_1$									
df$_2$	**Area**	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
20	**.10**	2.97	2.59	2.38	2.25	2.16	2.09	2.04	2.00	1.96	1.94
	.05	4.35	3.49	3.10	2.87	2.71	2.60	2.51	2.45	2.39	2.35
	.01	8.10	5.85	4.94	4.43	4.10	3.87	3.70	3.56	3.46	3.37
	.001	14.82	9.95	8.10	7.10	6.46	6.02	5.69	5.44	5.24	5.08
21	**.10**	2.96	2.57	2.36	2.23	2.14	2.08	2.02	1.98	1.95	1.92
	.05	4.32	3.47	3.07	2.84	2.68	2.57	2.49	2.42	2.37	2.32
	.01	8.02	5.78	4.87	4.37	4.04	3.81	3.64	3.51	3.40	3.31
	.001	14.59	9.77	7.94	6.95	6.32	5.88	5.56	5.31	5.11	4.95
22	**.10**	2.95	2.56	2.35	2.22	2.13	2.06	2.01	1.97	1.93	1.90
	.05	4.30	3.44	3.05	2.82	2.66	2.55	2.46	2.40	2.34	2.30
	.01	7.95	5.72	4.82	4.31	3.99	3.76	3.59	3.45	3.35	3.26
	.001	14.38	9.61	7.80	6.81	6.19	5.76	5.44	5.19	4.99	4.83
23	**.10**	2.94	2.55	2.34	2.21	2.11	2.05	1.99	1.95	1.92	1.89
	.05	4.28	3.42	3.03	2.80	2.64	2.53	2.44	2.37	2.32	2.27
	.01	7.88	5.66	4.76	4.26	3.94	3.71	3.54	3.41	3.30	3.21
	.001	14.20	9.47	7.67	6.70	6.08	5.65	5.33	5.09	4.89	4.73
24	**.10**	2.93	2.54	2.33	2.19	2.10	2.04	1.98	1.94	1.91	1.88
	.05	4.26	3.40	3.01	2.78	2.62	2.51	2.42	2.36	2.30	2.25
	.01	7.82	5.61	4.72	4.22	3.90	3.67	3.50	3.36	3.26	3.17
	.001	14.03	9.34	7.55	6.59	5.98	5.55	5.23	4.99	4.80	4.64
25	**.10**	2.92	2.53	2.32	2.18	2.09	2.02	1.97	1.93	1.89	1.87
	.05	4.24	3.39	2.99	2.76	2.60	2.49	2.40	2.34	2.28	2.24
	.01	7.77	5.57	4.68	4.18	3.85	3.63	3.46	3.32	3.22	3.13
	.001	13.88	9.22	7.45	6.49	5.89	5.46	5.15	4.91	4.71	4.56
26	**.10**	2.91	2.52	2.31	2.17	2.08	2.01	1.96	1.92	1.88	1.86
	.05	4.23	3.37	2.98	2.74	2.59	2.47	2.39	2.32	2.27	2.22
	.01	7.72	5.53	4.64	4.14	3.82	3.59	3.42	3.29	3.18	3.09
	.001	13.74	9.12	7.36	6.41	5.80	5.38	5.07	4.83	4.64	4.48
27	**.10**	2.90	2.51	2.30	2.17	2.07	2.00	1.95	1.91	1.87	1.85
	.05	4.21	3.35	2.96	2.73	2.57	2.46	2.37	2.31	2.25	2.20
	.01	7.68	5.49	4.60	4.11	3.78	3.56	3.39	3.26	3.15	3.06
	.001	13.61	9.02	7.27	6.33	5.73	5.31	5.00	4.76	4.57	4.41
28	**.10**	2.89	2.50	2.29	2.16	2.06	2.00	1.94	1.90	1.87	1.84
	.05	4.20	3.34	2.95	2.71	2.56	2.45	2.36	2.29	2.24	2.19
	.01	7.64	5.45	4.57	4.07	3.75	3.53	3.36	3.23	3.12	3.03
	.001	13.50	8.93	7.19	6.25	5.66	5.24	4.93	4.69	4.50	4.35
29	**.10**	2.89	2.50	2.28	2.15	2.06	1.99	1.93	1.89	1.86	1.83
	.05	4.18	3.33	2.93	2.70	2.55	2.43	2.35	2.28	2.22	2.18
	.01	7.60	5.42	4.54	4.04	3.73	3.50	3.33	3.20	3.09	3.00
	.001	13.39	8.85	7.12	6.19	5.59	5.18	4.87	4.64	4.45	4.29

(continued)

Table 6 Values That Capture Specified Upper-Tail *F* Curve Areas (Continued)

		df_1									
df_2	Area	1	2	3	4	5	6	7	8	9	10
30	.10	2.88	2.49	2.28	2.14	2.05	1.98	1.93	1.88	1.85	1.82
	.05	4.17	3.32	2.92	2.69	2.53	2.42	2.33	2.27	2.21	2.16
	.01	7.56	5.39	4.51	4.02	3.70	3.47	3.30	3.17	3.07	2.98
	.001	13.29	8.77	7.05	6.12	5.53	5.12	4.82	4.58	4.39	4.24
40	.10	2.84	2.44	2.23	2.09	2.00	1.93	1.87	1.83	1.79	1.76
	.05	4.08	3.23	2.84	2.61	2.45	2.34	2.25	2.18	2.12	2.08
	.01	7.31	5.18	4.31	3.83	3.51	3.29	3.12	2.99	2.89	2.80
	.001	12.61	8.25	6.59	5.70	5.13	4.73	4.44	4.21	4.02	3.87
60	.10	2.79	2.39	2.18	2.04	1.95	1.87	1.82	1.77	1.74	1.71
	.05	4.00	3.15	2.76	2.53	2.37	2.25	2.17	2.10	2.04	1.99
	.01	7.08	4.98	4.13	3.65	3.34	3.12	2.95	2.82	2.72	2.63
	.001	11.97	7.77	6.17	5.31	4.76	4.37	4.09	3.86	3.69	3.54
90	.10	2.76	2.36	2.15	2.01	1.91	1.84	1.78	1.74	1.70	1.67
	.05	3.95	3.10	2.71	2.47	2.32	2.20	2.11	2.04	1.99	1.94
	.01	6.93	4.85	4.01	3.53	3.23	3.01	2.84	2.72	2.61	2.52
	.001	11.57	7.47	5.91	5.06	4.53	4.15	3.87	3.65	3.48	3.34
120	.10	2.75	2.35	2.13	1.99	1.90	1.82	1.77	1.72	1.68	1.65
	.05	3.92	3.07	2.68	2.45	2.29	2.18	2.09	2.02	1.96	1.91
	.01	6.85	4.79	3.95	3.48	3.17	2.96	2.79	2.66	2.56	2.47
	.001	11.38	7.32	5.78	4.95	4.42	4.04	3.77	3.55	3.38	3.24
240	.10	2.73	2.32	2.10	1.97	1.87	1.80	1.74	1.70	1.65	1.63
	.05	3.88	3.03	2.64	2.41	2.25	2.14	2.04	1.98	1.92	1.87
	.01	6.74	4.69	3.86	3.40	3.09	2.88	2.71	2.59	2.48	2.40
	.001	11.10	7.11	5.60	4.78	4.25	3.89	3.62	3.41	3.24	3.09
∞	.10	2.71	2.30	2.08	1.94	1.85	1.77	1.72	1.67	1.63	1.60
	.05	3.84	3.00	2.60	2.37	2.21	2.10	2.01	1.94	1.88	1.83
	.01	6.63	4.61	3.78	3.32	3.02	2.80	2.64	2.51	2.41	2.32
	.001	10.83	6.91	5.42	4.62	4.10	3.74	3.47	3.27	3.10	2.96

Table 7 Critical Values of *q* for the Studentized Range Distribution

Error	Confidence	Number of populations, Treatments, or Levels being Compared							
df	Level	3	4	5	6	7	8	9	10
5	95%	4.60	5.22	5.67	6.03	6.33	6.58	6.80	6.99
	99%	6.98	7.80	8.42	8.91	9.32	9.67	9.97	10.24
6	95%	4.34	4.90	5.30	5.63	5.90	6.12	6.32	6.49
	99%	6.33	7.03	7.56	7.97	8.32	8.61	8.87	9.10
7	95%	4.16	4.68	5.06	5.36	5.61	5.82	6.00	6.16
	99%	5.92	6.54	7.01	7.37	7.68	7.94	8.17	8.37
8	95%	4.04	4.53	4.89	5.17	5.40	5.60	5.77	5.92
	99%	5.64	6.20	6.62	6.96	7.24	7.47	7.68	7.86
9	95%	3.95	4.41	4.76	5.02	5.24	5.43	5.59	5.74
	99%	5.43	5.96	6.35	6.66	6.91	7.13	7.33	7.49
10	95%	3.88	4.33	4.65	4.91	5.12	5.30	5.46	5.60
	99%	5.27	5.77	6.14	6.43	6.67	6.87	7.05	7.21
11	95%	3.82	4.26	4.57	4.82	5.03	5.20	5.35	5.49
	99%	5.15	5.62	5.97	6.25	6.48	6.67	6.84	6.99
12	95%	3.77	4.20	4.51	4.75	4.95	5.12	5.27	5.39
	99%	5.05	5.50	5.84	6.10	6.32	6.51	6.67	6.81
13	95%	3.73	4.15	4.45	4.69	4.88	5.05	5.19	5.32
	99%	4.96	5.40	5.73	5.98	6.19	6.37	6.53	6.67
14	95%	3.70	4.11	4.41	4.64	4.83	4.99	5.13	5.25
	99%	4.89	5.32	5.63	5.88	6.08	6.26	6.41	6.54
15	95%	3.67	4.08	4.37	4.59	4.78	4.94	5.08	5.20
	99%	4.84	5.25	5.56	5.80	5.99	6.16	6.31	6.44
16	95%	3.65	4.05	4.33	4.56	4.74	4.90	5.03	5.15
	99%	4.79	5.19	5.49	5.72	5.92	6.08	6.22	6.35
17	95%	3.63	4.02	4.30	4.52	4.70	4.86	4.99	5.11
	99%	4.74	5.14	5.43	5.66	5.85	6.01	6.15	6.27
18	95%	3.61	4.00	4.28	4.49	4.67	4.82	4.96	5.07
	99%	4.70	5.09	5.38	5.60	5.79	5.94	6.08	6.20
19	95%	3.59	3.98	4.25	4.47	4.65	4.79	4.92	5.04
	99%	4.67	5.05	5.33	5.55	5.73	5.89	6.02	6.14
20	95%	3.58	3.96	4.23	4.45	4.62	4.77	4.90	5.01
	99%	4.64	5.02	5.29	5.51	5.69	5.84	5.97	6.09
24	95%	3.53	3.90	4.17	4.37	4.54	4.68	4.81	4.92
	99%	4.55	4.91	5.17	5.37	5.54	5.69	5.81	5.92
30	95%	3.49	3.85	4.10	4.30	4.46	4.60	4.72	4.82
	99%	4.45	4.80	5.05	5.24	5.40	5.54	5.65	5.76
40	95%	3.44	3.79	4.04	4.23	4.39	4.52	4.63	4.73
	99%	4.37	4.70	4.93	5.11	5.26	5.39	5.50	5.60
60	95%	3.40	3.74	3.98	4.16	4.31	4.44	4.55	4.65
	99%	4.28	4.59	4.82	4.99	5.13	5.25	5.36	5.45
120	95%	3.36	3.68	3.92	4.10	4.24	4.36	4.47	4.56
	99%	4.20	4.50	4.71	4.87	5.01	5.12	5.21	5.30
∞	95%	3.31	3.63	3.86	4.03	4.17	4.29	4.39	4.47
	99%	4.12	4.40	4.60	4.76	4.88	4.99	5.08	5.16

Copyright 2025 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s). Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

Table 8 Upper-Tail Areas for Chi-Square Distributions

Right-tail area	**df = 1**	**df = 2**	**df = 3**	**df = 4**	**df = 5**
>0.100	<2.70	<4.60	<6.25	<7.77	<9.23
0.100	2.70	4.60	6.25	7.77	9.23
0.095	2.78	4.70	6.36	7.90	9.37
0.090	2.87	4.81	6.49	8.04	9.52
0.085	2.96	4.93	6.62	8.18	9.67
0.080	3.06	5.05	6.75	8.33	9.83
0.075	3.17	5.18	6.90	8.49	10.00
0.070	3.28	5.31	7.06	8.66	10.19
0.065	3.40	5.46	7.22	8.84	10.38
0.060	3.53	5.62	7.40	9.04	10.59
0.055	3.68	5.80	7.60	9.25	10.82
0.050	3.84	5.99	7.81	9.48	11.07
0.045	4.01	6.20	8.04	9.74	11.34
0.040	4.21	6.43	8.31	10.02	11.64
0.035	4.44	6.70	8.60	10.34	11.98
0.030	4.70	7.01	8.94	10.71	12.37
0.025	5.02	7.37	9.34	11.14	12.83
0.020	5.41	7.82	9.83	11.66	13.38
0.015	5.91	8.39	10.46	12.33	14.09
0.010	6.63	9.21	11.34	13.27	15.08
0.005	7.87	10.59	12.83	14.86	16.74
0.001	10.82	13.81	16.26	18.46	20.51
<0.001	>10.82	>13.81	>16.26	>18.46	>20.51

Right-tail area	**df = 6**	**df = 7**	**df = 8**	**df = 9**	**df = 10**
>0.100	<10.64	<12.01	<13.36	<14.68	<15.98
0.100	10.64	12.01	13.36	14.68	15.98
0.095	10.79	12.17	13.52	14.85	16.16
0.090	10.94	12.33	13.69	15.03	16.35
0.085	11.11	12.50	13.87	15.22	16.54
0.080	11.28	12.69	14.06	15.42	16.75
0.075	11.46	12.88	14.26	15.63	16.97
0.070	11.65	13.08	14.48	15.85	17.20
0.065	11.86	13.30	14.71	16.09	17.44
0.060	12.08	13.53	14.95	16.34	17.71
0.055	12.33	13.79	15.22	16.62	17.99
0.050	12.59	14.06	15.50	16.91	18.30
0.045	12.87	14.36	15.82	17.24	18.64
0.040	13.19	14.70	16.17	17.60	19.02
0.035	13.55	15.07	16.56	18.01	19.44
0.030	13.96	15.50	17.01	18.47	19.92
0.025	14.44	16.01	17.53	19.02	20.48
0.020	15.03	16.62	18.16	19.67	21.16
0.015	15.77	17.39	18.97	20.51	22.02
0.010	16.81	18.47	20.09	21.66	23.20
0.005	18.54	20.27	21.95	23.58	25.18
0.001	22.45	24.32	26.12	27.87	29.58
<0.001	>22.45	>24.32	>26.12	>27.87	>29.58

(continued)

Copyright 2025 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s). Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

Table 8 Upper-Tail Areas for Chi-Square Distributions (Continued)

Right-tail area	df = 11	df = 12	df = 13	df = 14	df = 15
>0.100	<17.27	<18.54	<19.81	<21.06	<22.30
0.100	17.27	18.54	19.81	21.06	22.30
0.095	17.45	18.74	20.00	21.26	22.51
0.090	17.65	18.93	20.21	21.47	22.73
0.085	17.85	19.14	20.42	21.69	22.95
0.080	18.06	19.36	20.65	21.93	23.19
0.075	18.29	19.60	20.89	22.17	23.45
0.070	18.53	19.84	21.15	22.44	23.72
0.065	18.78	20.11	21.42	22.71	24.00
0.060	19.06	20.39	21.71	23.01	24.31
0.055	19.35	20.69	22.02	23.33	24.63
0.050	19.67	21.02	22.36	23.68	24.99
0.045	20.02	21.38	22.73	24.06	25.38
0.040	20.41	21.78	23.14	24.48	25.81
0.035	20.84	22.23	23.60	24.95	26.29
0.030	21.34	22.74	24.12	25.49	26.84
0.025	21.92	23.33	24.73	26.11	27.48
0.020	22.61	24.05	25.47	26.87	28.25
0.015	23.50	24.96	26.40	27.82	29.23
0.010	24.72	26.21	27.68	29.14	30.57
0.005	26.75	28.29	29.81	31.31	32.80
0.001	31.26	32.90	34.52	36.12	37.69
<0.001	>31.26	>32.90	>34.52	>36.12	>37.69

Right-tail area	df = 16	df = 17	df = 18	df = 19	df = 20
>0.100	<23.54	<24.77	<25.98	<27.20	<28.41
0.100	23.54	24.77	25.98	27.20	28.41
0.095	23.75	24.98	26.21	27.43	28.64
0.090	23.97	25.21	26.44	27.66	28.88
0.085	24.21	25.45	26.68	27.91	29.14
0.080	24.45	25.70	26.94	28.18	29.40
0.075	24.71	25.97	27.21	28.45	29.69
0.070	24.99	26.25	27.50	28.75	29.99
0.065	25.28	26.55	27.81	29.06	30.30
0.060	25.59	26.87	28.13	29.39	30.64
0.055	25.93	27.21	28.48	29.75	31.01
0.050	26.29	27.58	28.86	30.14	31.41
0.045	26.69	27.99	29.28	30.56	31.84
0.040	27.13	28.44	29.74	31.03	32.32
0.035	27.62	28.94	30.25	31.56	32.85
0.030	28.19	29.52	30.84	32.15	33.46
0.025	28.84	30.19	31.52	32.85	34.16
0.020	29.63	30.99	32.34	33.68	35.01
0.015	30.62	32.01	33.38	34.74	36.09
0.010	32.00	33.40	34.80	36.19	37.56
0.005	34.26	35.71	37.15	38.58	39.99
0.001	39.25	40.78	42.31	43.81	45.31
<0.001	>39.25	>40.78	>42.31	>43.81	>45.31

Table 9 Binomial Probabilities

$n = 5$

	p												
x	0.05	0.1	0.2	0.25	0.3	0.4	0.5	0.6	0.7	0.75	0.8	0.9	0.95
0	.774	.590	.328	.237	.168	.078	.031	.010	.002	.001	.000	.000	.000
1	.204	.328	.410	.396	.360	.259	.156	.077	.028	.015	.006	.000	.000
2	.021	.073	.205	.264	.309	.346	.313	.230	.132	.088	.051	.008	.001
3	.001	.008	.051	.088	.132	.230	.313	.346	.309	.264	.205	.073	.021
4	.000	.000	.006	.015	.028	.077	.156	.259	.360	.396	.410	.328	.204
5	.000	.000	.000	.001	.002	.010	.031	.078	.168	.237	.328	.590	.774

$n = 10$

	p												
x	0.05	0.1	0.2	0.25	0.3	0.4	0.5	0.6	0.7	0.75	0.8	0.9	0.95
0	.599	.349	.107	.056	.028	.006	.001	.000	.000	.000	.000	.000	.000
1	.315	.387	.268	.188	.121	.040	.010	.002	.000	.000	.000	.000	.000
2	.075	.194	.302	.282	.233	.121	.044	.011	.001	.000	.000	.000	.000
3	.010	.057	.201	.250	.267	.215	.117	.042	.009	.003	.001	.000	.000
4	.001	.011	.088	.146	.200	.251	.205	.111	.037	.016	.006	.000	.000
5	.000	.001	.026	.058	.103	.201	.246	.201	.103	.058	.026	.001	.000
6	.000	.000	.006	.016	.037	.111	.205	.251	.200	.146	.088	.011	.001
7	.000	.000	.001	.003	.009	.042	.117	.215	.267	.250	.201	.057	.010
8	.000	.000	.000	.000	.001	.011	.044	.121	.233	.282	.302	.194	.075
9	.000	.000	.000	.000	.000	.002	.010	.040	.121	.188	.268	.387	.315
10	.000	.000	.000	.000	.000	.000	.001	.006	.028	.056	.107	.349	.599

(continued)

Table 9 Binomial Probabilities (Continued)

***n* = 15**

	p												
x	**0.05**	**0.1**	**0.2**	**0.25**	**0.3**	**0.4**	**0.5**	**0.6**	**0.7**	**0.75**	**0.8**	**0.9**	**0.95**
0	.463	.206	.035	.013	.005	.000	.000	.000	.000	.000	.000	.000	.000
1	.366	.343	.132	.067	.031	.005	.000	.000	.000	.000	.000	.000	.000
2	.135	.267	.231	.156	.092	.022	.003	.000	.000	.000	.000	.000	.000
3	.031	.129	.250	.225	.170	.063	.014	.002	.000	.000	.000	.000	.000
4	.005	.043	.188	.225	.219	.127	.042	.007	.001	.000	.000	.000	.000
5	.001	.010	.103	.165	.206	.186	.092	.024	.003	.001	.000	.000	.000
6	.000	.002	.043	.092	.147	.207	.153	.061	.012	.003	.001	.000	.000
7	.000	.000	.014	.039	.081	.177	.196	.118	.035	.013	.003	.000	.000
8	.000	.000	.003	.013	.035	.118	.196	.177	.081	.039	.014	.000	.000
9	.000	.000	.001	.003	.012	.061	.153	.207	.147	.092	.043	.002	.000
10	.000	.000	.000	.001	.003	.024	.092	.186	.206	.165	.103	.010	.001
11	.000	.000	.000	.000	.001	.007	.042	.127	.219	.225	.188	.043	.005
12	.000	.000	.000	.000	.000	.002	.014	.063	.170	.225	.250	.129	.031
13	.000	.000	.000	.000	.000	.000	.003	.022	.092	.156	.231	.267	.135
14	.000	.000	.000	.000	.000	.000	.000	.005	.031	.067	.132	.343	.366
15	.000	.000	.000	.000	.000	.000	.000	.000	.005	.013	.035	.206	.463

***n* = 20**

	p												
x	**0.05**	**0.1**	**0.2**	**0.25**	**0.3**	**0.4**	**0.5**	**0.6**	**0.7**	**0.75**	**0.8**	**0.9**	**0.95**
0	.358	.122	.012	.003	.001	.000	.000	.000	.000	.000	.000	.000	.000
1	.377	.270	.058	.021	.007	.000	.000	.000	.000	.000	.000	.000	.000
2	.189	.285	.137	.067	.028	.003	.000	.000	.000	.000	.000	.000	.000
3	.060	.190	.205	.134	.072	.012	.001	.000	.000	.000	.000	.000	.000
4	.013	.090	.218	.190	.130	.035	.005	.000	.000	.000	.000	.000	.000
5	.002	.032	.175	.202	.179	.075	.015	.001	.000	.000	.000	.000	.000
6	.000	.009	.109	.169	.192	.124	.037	.005	.000	.000	.000	.000	.000
7	.000	.002	.055	.112	.164	.166	.074	.015	.001	.000	.000	.000	.000
8	.000	.000	.022	.061	.114	.180	.120	.035	.004	.001	.000	.000	.000
9	.000	.000	.007	.027	.065	.160	.160	.071	.012	.003	.000	.000	.000
10	.000	.000	.002	.010	.031	.117	.176	.117	.031	.010	.002	.000	.000
11	.000	.000	.000	.003	.012	.071	.160	.160	.065	.027	.007	.000	.000
12	.000	.000	.000	.001	.004	.035	.120	.180	.114	.061	.022	.000	.000
13	.000	.000	.000	.000	.001	.015	.074	.166	.164	.112	.055	.002	.000
14	.000	.000	.000	.000	.000	.005	.037	.124	.192	.169	.109	.009	.000
15	.000	.000	.000	.000	.000	.001	.015	.075	.179	.202	.175	.032	.002
16	.000	.000	.000	.000	.000	.000	.005	.035	.130	.190	.218	.090	.013
17	.000	.000	.000	.000	.000	.000	.001	.012	.072	.134	.205	.190	.060
18	.000	.000	.000	.000	.000	.000	.000	.003	.028	.067	.137	.285	.189
19	.000	.000	.000	.000	.000	.000	.000	.000	.007	.021	.058	.270	.377
20	.000	.000	.000	.000	.000	.000	.000	.000	.001	.003	.012	.122	.358

(continued)

Table 9 **Binomial Probabilities (Continued)**

***n* = 25**

	p												
x	**0.05**	**0.1**	**0.2**	**0.25**	**0.3**	**0.4**	**0.5**	**0.6**	**0.7**	**0.75**	**0.8**	**0.9**	**0.95**
0	.277	.072	.004	.001	.000	.000	.000	.000	.000	.000	.000	.000	.000
1	.365	.199	.024	.006	.001	.000	.000	.000	.000	.000	.000	.000	.000
2	.231	.266	.071	.025	.007	.000	.000	.000	.000	.000	.000	.000	.000
3	.093	.226	.136	.064	.024	.002	.000	.000	.000	.000	.000	.000	.000
4	.027	.138	.187	.118	.057	.007	.000	.000	.000	.000	.000	.000	.000
5	.006	.065	.196	.165	.103	.020	.002	.000	.000	.000	.000	.000	.000
6	.001	.024	.163	.183	.147	.044	.005	.000	.000	.000	.000	.000	.000
7	.000	.007	.111	.165	.171	.080	.014	.001	.000	.000	.000	.000	.000
8	.000	.002	.062	.124	.165	.120	.032	.003	.000	.000	.000	.000	.000
9	.000	.000	.029	.078	.134	.151	.061	.009	.000	.000	.000	.000	.000
10	.000	.000	.012	.042	.092	.161	.097	.021	.001	.000	.000	.000	.000
11	.000	.000	.004	.019	.054	.147	.133	.043	.004	.001	.000	.000	.000
12	.000	.000	.001	.007	.027	.114	.155	.076	.011	.002	.000	.000	.000
13	.000	.000	.000	.002	.011	.076	.155	.114	.027	.007	.001	.000	.000
14	.000	.000	.000	.001	.004	.043	.133	.147	.054	.019	.004	.000	.000
15	.000	.000	.000	.000	.001	.021	.097	.161	.092	.042	.012	.000	.000
16	.000	.000	.000	.000	.000	.009	.061	.151	.134	.078	.029	.000	.000
17	.000	.000	.000	.000	.000	.003	.032	.120	.165	.124	.062	.002	.000
18	.000	.000	.000	.000	.000	.001	.014	.080	.171	.165	.111	.007	.000
19	.000	.000	.000	.000	.000	.000	.005	.044	.147	.183	.163	.024	.001
20	.000	.000	.000	.000	.000	.000	.002	.020	.103	.165	.196	.065	.006
21	.000	.000	.000	.000	.000	.000	.000	.007	.057	.118	.187	.138	.027
22	.000	.000	.000	.000	.000	.000	.000	.002	.024	.064	.136	.226	.093
23	.000	.000	.000	.000	.000	.000	.000	.000	.007	.025	.071	.266	.231
24	.000	.000	.000	.000	.000	.000	.000	.000	.002	.006	.024	.199	.365
25	.000	.000	.000	.000	.000	.000	.000	.000	.000	.001	.004	.072	.277

Answers to Selected Odd-Numbered Exercises

Chapter 1

1.1 *Descriptive statistics* is the branch of statistics that involves the organization and summary of the values in a data set. *Inferential statistics* is the branch of statistics concerned with reaching conclusions about a population based on the information provided by a sample.

1.3 **a.** census **b.** population value

1.5 **a.** The population of interest is the set of all 15,000 students at the university. **b.** The sample is the 200 students who are interviewed.

1.7 The population is the set of all 7000 property owners. The sample is the group of 500 owners included in the survey.

1.9 The population is the set of 5000 used bricks. The sample is the set of 100 bricks they check.

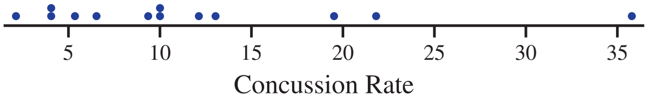
1.11 **a.** The researchers were trying to learn about the effect of blue light exposure on the quality and ripeness of mangoes. The motivation for their research could include wanting to be able to sell better quality and speed up the ripening process of mangoes. **b.** The study was conducted in a reasonable way because the mangoes that were exposed to blue light were compared to mangoes that were kept in the dark. This way, one can be confident that differences can be attributed the presence or absence of blue light. I would want to know if the mangoes in the two groups were kept at the same temperature and were picked at the same time.

1.13 **a.** Categorical **b.** Categorical **c.** Numerical (discrete) **d.** Numerical (continuous) **e.** Categorical **f.** Numerical (continuous)

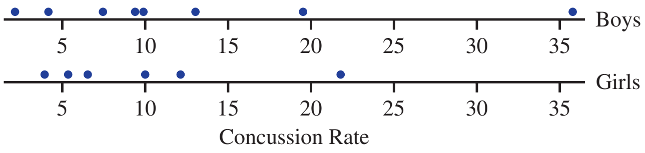
1.15 **a.** Continuous **b.** Continuous **c.** Continuous **d.** Discrete

1.17 **a.** Color, brand of motorcycle, telephone area code **b.** Number of previous motorcycles **c.** Bar chart **d.** Dotplot

1.19 **a.**

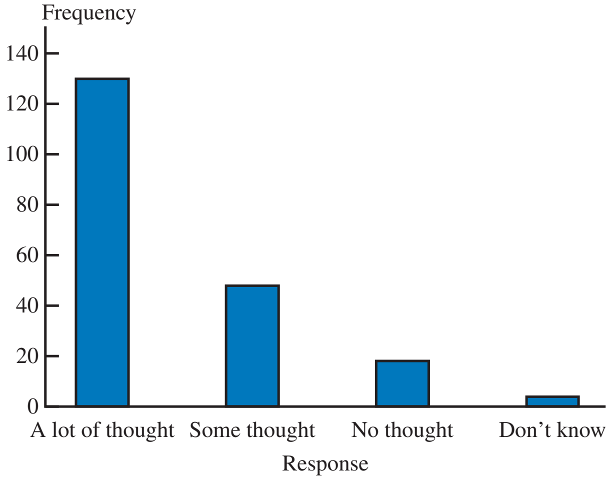


b.



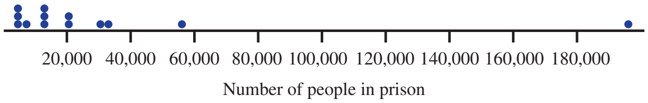
The sport with an unusually high (compared to all the other sports) concussion rate is football. Without considering football, the concussion rate for girls' sports is essentially the same as the concussion rate for boys' sports. However, if we consider football, the concussion rate for girls' sports tends to be lower than that for boys' sports.

1.21 **a.**



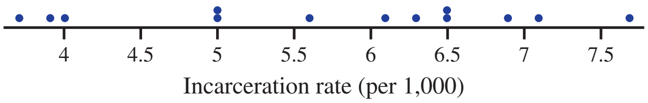
b. The most common response was "A lot of thought," accounting for 130 (or 65%) of the students who started college but did not complete a degree. The next two most common responses were "Some thought" and "No thought," accounting for 48 (or 24%) and 18 (or 9%), respectively, of the students who started college but did not complete a degree. Finally, 4 of the 200 respondents (2%) indicated that don't know how much thought they have given to going back to school.

1.23 **a.**



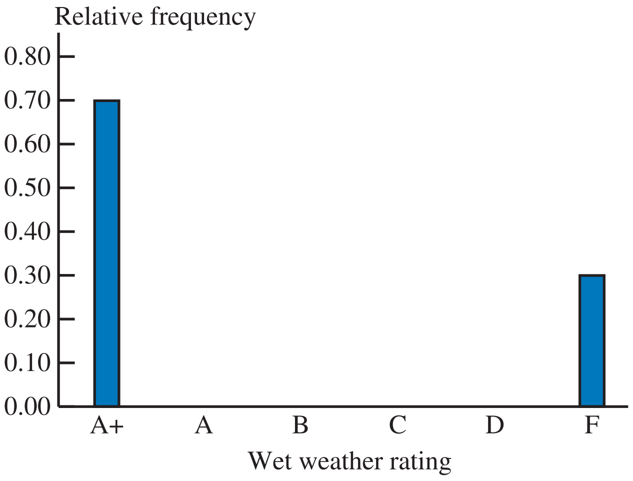
California, with 196,100 people in prison, stands out from the remaining Western states. Additionally, Arizona stands out as well with 56,000 people in prison.
b. The incarceration rates for all the Western states are shown in the table below, followed by the corresponding dotplot.

State	Incarceration Rate
Alaska	6.1
Arizona	7.7
California	5.0
Colorado	5.6
Hawaii	3.7
Idaho	7.1
Montana	6.5
Nevada	6.5
New Mexico	6.3
Oregon	5.0
Utah	3.9
Washington	4.0
Wyoming	6.9



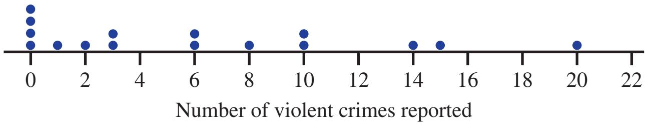
No, the same states do not stand out as unusual. Now only Arizona is somewhat unusual on the high end, and three states are unusual on the low end (Hawaii, Utah, and Washington).
c. A typical number of people in prison is 13,300. The dotplot shows that there are two large values (California with 196,100 and Arizona with 56,000). The other Western states have fewer people in prison. Overall, the number of people in prison ranged between 4,000 and 196,100. Considering the incarceration rates (per 1,000), a typical value is 6.1, with values that range between 3.7 and 7.7 people in prison per 1,000.

1.25 a.

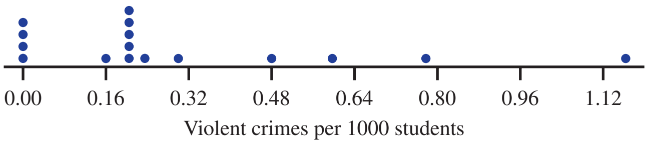


b. Seventy-five percent (75%) of the dry weather ratings are B or higher, and 70% of wet weather ratings are B or higher, indicating that dry weather ratings are higher than wet weather ratings. Note that the wet weather ratings are only A+ or F, so the wet weather ratings are more extreme than dry weather ratings.

1.27 a.

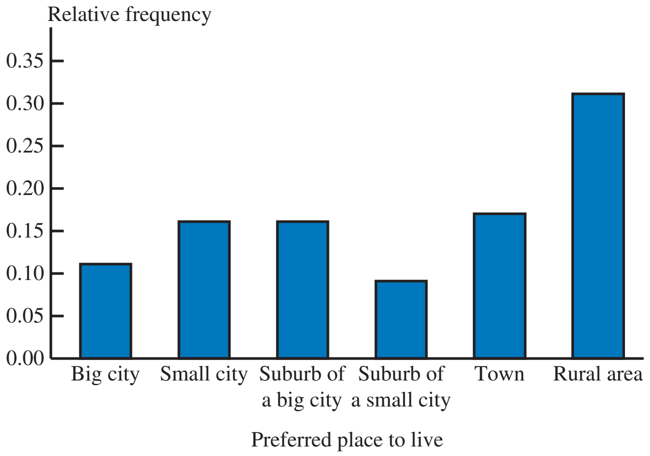


Three schools seem to stand out from the rest (in increasing order of number of crimes): University of Central Florida (14 crimes reported), Florida International University (15 crimes reported), and Florida State University (20 crimes reported).
b.



The colleges that stand out in violent crimes per 1000 students are (in increasing order of crime rate) Florida State University, Florida A&M University, University of West Florida, and New College of Florida. Only Florida State University stands out in both dotplots. **c.** For the number of violent crimes, there are three schools that stand out by having high numbers of crimes, with the majority of the schools having similar, and low (10 or fewer), numbers of crimes. There seems to be greater consistency for crime rate (per 1000 students) among the 16 schools than there is for number of crimes, with four schools standing out as having high crime rates, and four schools with crime rates that stand out as being noticeably low.

1.29 a.



b. Of the survey respondents, 31% indicated that their preferred place to live is a rural area, which is the highest percentage of the categories. Seventeen percent (17%) of respondents indicated that a town is their preferred place

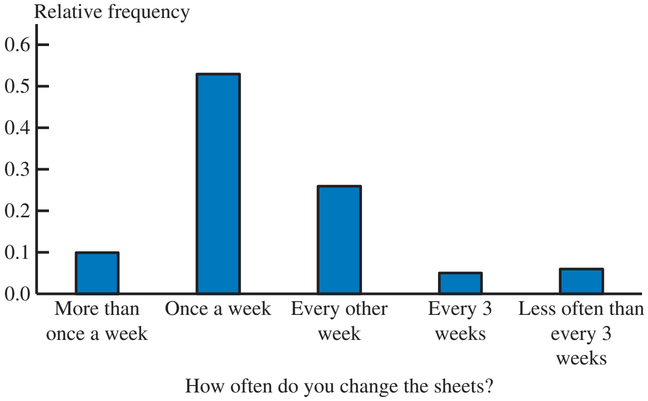
Copyright 2025 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s). Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

to live, followed closely by small city and suburb of a big city, each with 16% of the respondents indicating that is their preferred place to live. Big city and suburb of small city received the lowest percentages of respondents, with 11% and 9%, respectively.

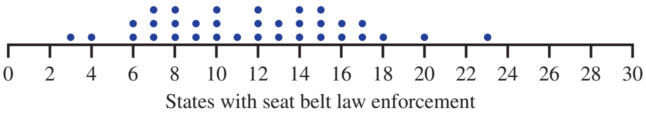
1.31 **a.**

How Often	Relative Frequency
More than once a week	0.10
Once a week	0.53
Every other week	0.26
Every 3 weeks	0.05
Less often than every 3 weeks	0.06

b.

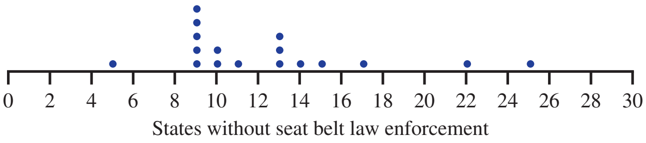


1.33 **a.**



The distribution of fatality rates for states with seat belt law enforcement is approximately symmetric. The typical value is approximately 12 traffic deaths per 100,000 population. The fatality rates vary between approximately 3 and 23 traffic deaths per 100,000 population.

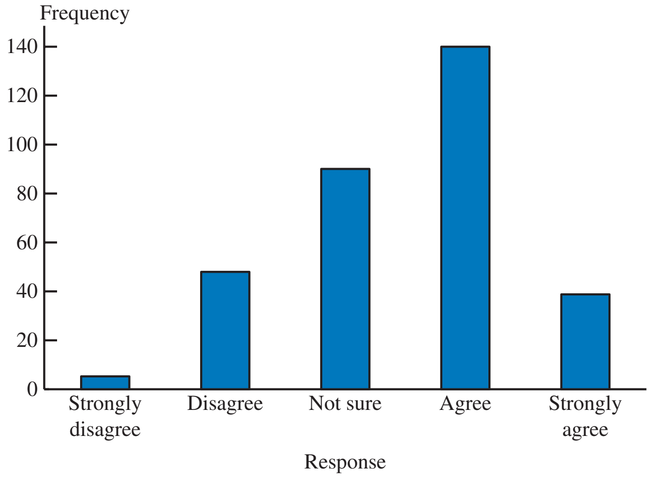
b.



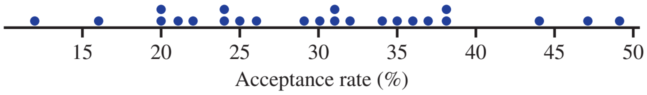
The distribution of fatality rates for states without seat belt law enforcement is asymmetric, with more values clustered below the typical value of approximately 11 traffic deaths per 100,000 population. The fatality rates vary between approximately 5 and 25 traffic deaths per 100,000 population. There are two values that are unusually large (22 and 25), and one value that is rather small (5).

c. The centers of the distributions are approximately the same (11 and 12), and the variability of the two distributions are approximately the same (both distributions have values that span approximately 20 traffic deaths per 100,000 population). The distribution of fatality rates for states with seat belt laws is roughly symmetric, whereas the distribution of fatality rates for states without seat belt laws is asymmetric.

1.35



1.37 **a.**



b. A typical acceptance rate for these top 25 schools is around 30, with the great majority of acceptance rates being between 19 and 39. There are no particularly extreme values. The pattern of the points is roughly symmetrical.

Chapter 2

2.1 Observational study, because the researchers used data from hospital emergency departments and did not assign participants to the dog walking groups.

2.3 Experiment because the course professor assigned each discussion group an instructor who taught under a particular gender identity (perceived gender was the experimental treatment).

2.5 **a.** Experiment, because researchers decided which participants would receive which treatments. **b.** Yes, because the participants were randomly assigned to the treatments.

2.7 **a.** Experiment. **b.** Yes, because the participants were randomly assigned to the treatments.

2.9 We would need to know if the teens were randomly selected from the population of teens who spend time on a computer. In order to generalize the given conclusion to all teens who spend time on a computer, a random sample from the population of interest must be taken.

2.11 **a.** The data would need to be collected from a simple random sample of all adult Americans. **b.** No. Since the survey was conducted by telephone, those adult Americans without telephones would be excluded from the sample, and not all adult Americans would answer the telephone.

Therefore, we can only generalize to the population of adult Americans who have telephones and would answer their phone.

2.13 Number the names on the list from 1 to *n*, where *n* is the number of students at the college. Then use a random number generator to select 100 different numbers between 1 and *n*. Students corresponding to these 100 numbers would be included in the sample.

2.15 Number the names on list, use random number generator to select 30 numbers between 1 and 500. Signatures corresponding to the 30 selected numbers constitute the random sample.

2.17 **a.** The population of interest is all Arizona drivers. **b.** Although the details of the sampling scheme are not presented, there was random selection of AAA Arizona members. Therefore, the sample was selected in a reasonable way. **c.** This sample is not representative of the population of interest because only those people who are AAA Arizona members were included in the sample. **d.** Selection bias is present because those drivers who are not AAA Arizona members have been excluded from the sample. Nonresponse bias could also be present because there was no indication of how many surveys were not returned.

2.19 **a.** Using the list, first number the part-time students 1–3000. Use a random number generator on a calculator or computer to randomly select a whole number between 1 and 3000. The number selected represents the first part-time student to be included in the sample. Repeat the number selection, ignoring repeated numbers, until 10 part-time students have been selected. Then number the full-time students 1–3500 and select 10 full-time students using the same procedure. **b.** No

2.21 **a.** The pages of the book have already been numbered between 1 and the highest page number in the book. Use a random number generator on a calculator or computer to randomly select a whole number between 1 and the highest page number in the book. The number selected will be the first page to be included in the sample. Repeat the number selection, ignoring repeated numbers, until the required number of pages has been selected. **b.** Pages that include exercises tend to contain more words than pages that do not include exercises. Therefore, it would be sensible to stratify according to this criterion. Assuming that 20 nonexercise pages and 20 exercise pages will be included in the sample, the sample should be selected as follows: Use a random number generator to randomly select a whole number between 1 and the highest page number in the book; the number selected will be the first page to be included in the sample; repeat the number selection, ignoring repeated numbers and keeping track of the number of pages of each type selected, until 20 pages of one type have been selected; then continue in the same way, but ignore numbers corresponding to pages of that type; when 20 pages of the other type have been selected, stop the process. **c.** Randomly select one page from the first 20 pages in the book. Include in your sample that page and every 20th page from that page onward. **d.** Roughly speaking, in terms of the numbers of words per page, each chapter is representative of the book as a whole. It is therefore sensible for the chapters to be used as clusters. Using a random number generator, randomly choose three chapters. Then count the number of words on each page in those three chapters.

2.23 Answers will vary.

2.25 **a.** The population of interest is all college students. **b.** If the researchers want to generalize to all college students, this sample might not be reasonable. For example, although there are students from different majors in the class, it might be that not all majors are represented. Additionally, giving extra credit encourages those students who need the extra credit or those who are interested in the study to participate; those students who don't need the extra credit or those who don't care much might not participate. There might be systematic differences between those who need the extra credit and those who don't. **c.** Given that only students in this class were surveyed, the sample is not likely to be representative of the population of interest (all college students). **d.** Selection bias is present because the selection method might exclude students in certain majors or those students who don't need the extra credit.

2.27 It is not reasonable to consider the participants to be representative of all students with regard to their truthfulness in the various forms of communication. Also, the students knew they were surveying themselves as to the truthfulness of their interactions. This could easily have changed their behavior in particular social contexts and, therefore, could have distorted the results of the study.

2.29 One factor that might contribute to these very different results has to do with who uses landlines. Perhaps older Alabama residents, who might be more conservative and Republican-leaning than younger Alabama residents, have landlines. Younger and perhaps somewhat less conservative Alabama residents might exclusively use cell phones.

2.31 **a.** Yes. It is possible that students of a given class-standing tend to be similar in the amounts of money they spend on textbooks. **b.** Yes. It is possible that students who pursue a certain field of study tend to be similar in the amounts of money they spend on textbooks. **c.** No. It is unlikely that stratifying in this way will produce groups that are homogeneous in terms of the students' spending on textbooks.

2.33 No, because the people who sent their hair for testing did so voluntarily. It is quite possible that people who would choose to participate in a study of this sort differ in their mercury levels from the population as a whole.

2.35 **a.** Binding strength **b.** Type of glue **c.** The extraneous variables mentioned are the number of pages in the book and whether the book is bound as a hardback or a paperback. Further extraneous variables that might be

Copyright 2025 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s). Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

considered include the weight of the material used for the cover and the type of paper used.

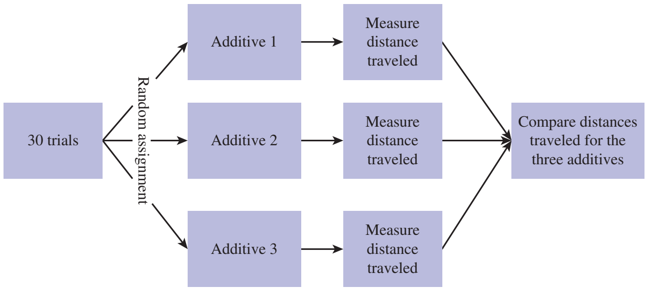
2.37 Answers will vary. One possible answer: Write the names of the 30 subjects on slips of paper. Mix the slips and then select 10. Assign these 10 subjects to hand-drying method 1. Mix the remaining slips and then select 10. Assign these 10 subjects to hand-drying method 2. Assign the remaining 10 subjects to method 3.

2.39 **a.** Blocking **b.** Direct control

2.41 The figure shows that comparable groups in terms of age have been formed.

2.43 **a.** The experimental conditions were the presence or absence of music with a vocal component. The response is the time required to complete the surgical procedure. **b.** Some surgical procedures are more complex and require a greater degree of concentration; music with a vocal component might be more distracting when the surgical procedure is more complex. The temperature of the room might affect the comfort of the surgeon; if the surgeon is too hot or too cold, they might be uncomfortable, and therefore more easily distracted by the vocal component. If the music is too loud, the surgeon might be distracted and unable to focus, regardless of the presence or absence of the vocal component. If the music is too soft, the surgeon might try to concentrate on listening to the vocal component, and therefore pay more attention to the music rather than the surgical procedure. **c.** Random assignment of surgeons to music condition is important because there might be something inherently different about surgeons who want no vocals versus those who do want vocals. Random assignment ensures that the experiment does not favor one experimental condition over another.

2.45


30 trials
Random assignment
Additive 1
Additive 2
Additive 3
Measure distance traveled
Measure distance traveled
Measure distance traveled
Compare distances traveled for the three additives

2.47 **a.** Experiment **b.** No **c.** Yes **d.** Yes, because the experiment used random assignment of subjects to treatments. **e.** No, because the subjects were not randomly selected.

2.49 **a.** The treatments are wearing glasses and not wearing glasses in public spaces. **b.** The response variable is whether or not the subject tested positive for COVID-19.

2.51 In order to determine if the conclusions implied by the headline are appropriate, you would need to know if dieters were randomly assigned to the experimental conditions (large fork or small fork). In order to generalize to the population of dieters, you would also want to know if the study participants were randomly selected from the population of dieters.

2.53 "Blinding" is ensuring that the experimental subjects do not know which treatment they were given and/or ensuring that the people who measure the response variable do not know who was given which treatment.

2.55 **a.** Allowing study participants to choose which group they want to be in could introduce systematic differences between the two experimental conditions (knee replacement surgery with exercise and exercise therapy alone), resulting in potential confounding. Those who chose knee replacement surgery plus exercise might, in some way, be different from those who chose exercise therapy alone. We would not know if differences in pain relief between the two groups were due to the knee replacement surgery with exercise or due to some inherent differences in the subjects who chose their experimental groups. **b.** The researchers likely did not include a control group because the study participants needed some relief from their pain. Because the purpose of this experiment is to determine whether knee replacement surgery with exercise provided more pain relief than exercise therapy alone, a control group would not allow the study participants to have the opportunity to experience pain relief.

2.57 We will assume that only four colors will be compared, and that only headache sufferers will be included in the study. Prepare a supply of "Regular Strength" Tylenol in four different colors: white (the current color of the medication, and therefore the "control"), red, green, and blue. Recruit 20 volunteers who suffer from headaches. Instruct each volunteer not to take any pain relief medication for a week. After that week is over, issue each volunteer a supply of all four colors. Give each volunteer an order in which to use the colors (this order would be determined randomly for each volunteer). Instruct the volunteers to use one fixed dose of the medication for each headache over a period of 4 weeks, and to note on a form the color used and the pain relief achieved (on a scale of 0–10, where 0 is no pain relief and 10 is complete pain relief). At the end of the 4 weeks gather the results and compare the pain relief achieved by the four colors.

2.59 Answers will vary. One possible answer: Of the children who were in first grade, randomly assign 350 to the book group and 350 to the no-book group. You could do this by starting with a computer list of the names of the 700 first-graders. Assign to each name a random 10-digit number. Sort the names according to the numbers, from smallest to largest. The first 350 names on the sorted list are assigned to the book group, and the remainder to the no-book group. Using a similar method, randomly assign 315 of the second-graders to the book group and 315 to the no-book group.

2.61 **a.** If the judges had known which chowder came from which restaurant, then it is unlikely that Denny's chowder would have won the contest, since the judges would probably be conditioned by this knowledge to choose chowders from more expensive restaurants. **b.** In experiments, if the people measuring the response are not blinded, they will often be conditioned to see different responses to some treatments over other treatments, in the same way as the judges would have been conditioned to favor the expensive restaurant chowders. Therefore, it is necessary that the people measuring the response should not know which subject received which treatment, so that the treatments can be compared on their own merits.

2.63 **a.** A placebo group would be necessary if the mere thought of having amalgam fillings could produce kidney disorders. However, since the experimental subjects were sheep, the researchers do not need to be concerned that this would happen. **b.** A resin filling treatment group would be necessary in order to provide evidence that it is the material in the amalgam fillings, rather than the process of filling the teeth, or just the presence of foreign bodies in the teeth, that is the cause of the kidney disorders. If the amalgam filling group developed the kidney disorders and the resin filling group did not, then this would provide evidence that it is some ingredient in the amalgam fillings that is causing the kidney problems. **c.** Since there is concern about the effect of amalgam fillings, it would be considered unethical to use humans in the experiment.

2.65 **a.** This is an observational study. **b.** In order to evaluate the study, we need to know whether the sample was a random sample.

2.67 Answers will vary.

2.69 Answers will vary.

2.71 It is not reasonable to conclude that a Southern-style diet causes higher rates of cardiac death because this was an observational study and cause-and-effect conclusions cannot be drawn.

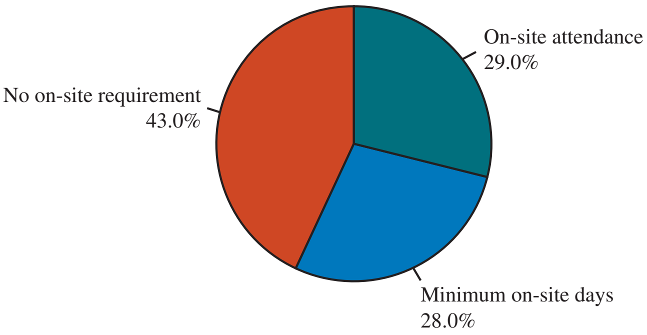
2.73 A control group (in this case, having a group of study participants take the test at the same time intervals without having consumed any THC) would allow the experimenter to assess how the response variable behaves when the treatment is not used. It would provide a baseline against which of the treatment groups can be compared to determine the effect that THC has on driving. The fact that there is no random assignment of THC levels to trials doesn't allow for comparison of different THC levels.

2.75 This is an observational study based on results of a survey (no consumers were assigned to different experimental conditions).

2.77 **a.** The design is good in that it includes several desired experimental design components. Specifically, the design includes a control group, random assignment of the subjects to the treatments, and includes blinding.
b. Rather than taking photos of the tops of the heads of all the women, the expert who determined the change in hair density should have the opportunity to evaluate the women in person. Additionally, there could have been more than one expert doing the change in hair density evaluation.

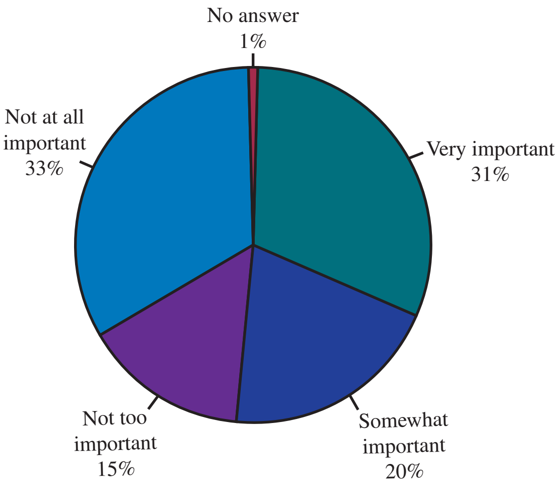
Chapter 3

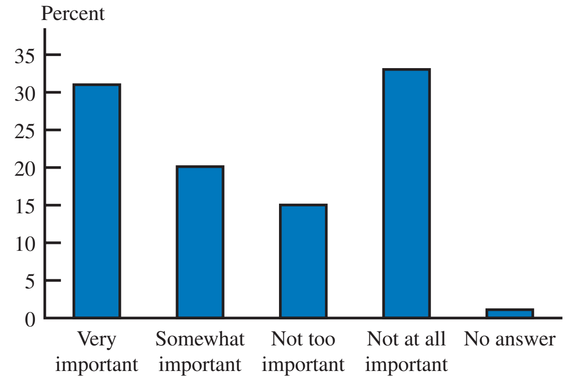
3.1



3.3 Answers will vary. One notable difference is that the percent of teens who use Instagram and Snapchat increased noticeably between 2014–2015 and 2022. However, there was a large drop in the percent of teens who use Facebook during that same time period, and smaller but noticeable drops in the percent of teens who use Twitter and Tumblr.

3.5 **a.**



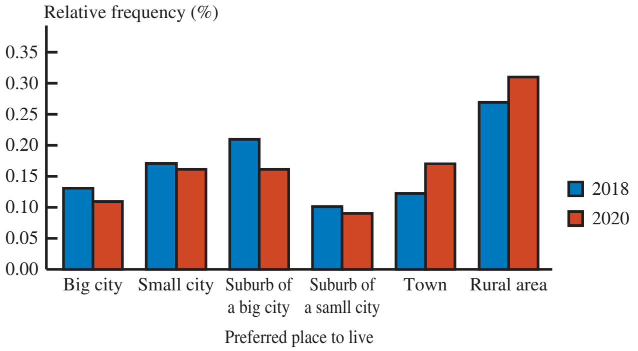


b. The bar chart is more effective for communicating the survey results because the differences in percents (represented by bar height) are easier to see and compare than the sizes of the pie pieces.

3.7 **a.** This graph is similar to a pie chart in that the areas of the labeled regions are proportional to the percentages for each category, and the areas sum to 100%. **b.** A pie chart would not have been as effective in summarizing how the federal government spends money because there are so many categories and the pie chart would be difficult to read.

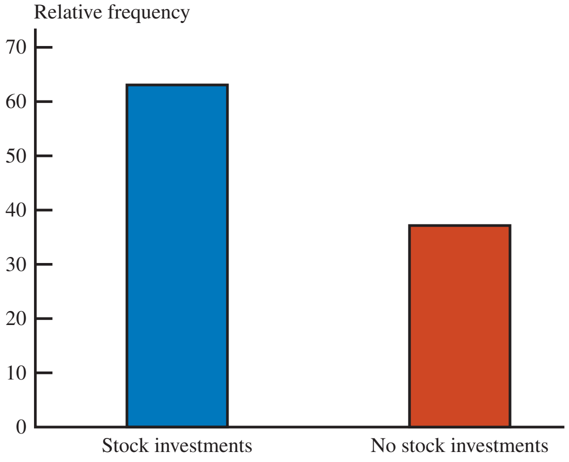
3.9 **a.** This graph is similar to a pie chart because the portions of the house add up to 100%, as they should in a pie chart. **b.** Chart III is the segmented bar chart that is a graph of the data used to create the USA Today Snapshot graph. The height of the "Yes" portion of the bar in Chart III has a height that corresponds to 23%. The "No" portion of the bar in Chart III has a height that corresponds to 48%, which gives a total height of 71% for the combined "Yes" and "No" responses. Finally, the "Not sure" portion of the bar makes up the remaining 29%, making the total 100%.

3.11 **a.**

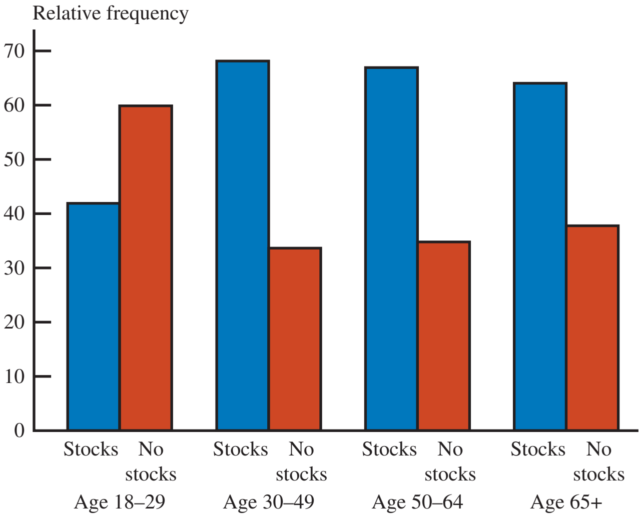


b. When comparing the results from 2020 and 2018, rural area is the most frequently chosen preferred place to live, selected by 31% of the respondents in 2020 and 27% in 2018. Similarly, the least frequently chosen preferred place to live is suburb of a small city, selected by 9% of the respondents in 2020 and 10% of the respondents in 2018. Notice as well that big city, small city, suburb of big city, and suburb of small city all saw a decrease between 2018 and 2020, whereas town and rural area both saw an increase between 2018 and 2020. **c.** The conclusion that "Country Living Enjoys Renewed Appeal in U.S." is supported by the comparative bar chart in that town and rural areas saw an increase in respondents between 2018 and 2020.

3.13 **a.**



b.



3.15

9	345
10	0457
11	01456889
12	00022333334445666677899
13	124689
14	12479
15	3
16	9

Stem: Ones
Leaf: Tenths

A typical number of births per thousand of the population is around 12.4, with most birth rates concentrated in the 11.0 to 13.9 range. The distribution has just one peak (at the 12–13 class). There are no extreme values. The distribution is approximately symmetrical.

3.17 **a.**

31	0
32	
33	
34	0 5 7
35	4
36	3 5 5 6 7 8 8 8 9
37	5 5 7 7 7 7 9
38	0 0 1 2 3 5 6 7 7 8 9 9
39	2 3 3 4 5 5 5 8 8
40	0 1 8 9
41	
42	0 4 6
43	2
44	6

Median age

KEY: 44|6 = 44.6

The distribution of median ages is centered at approximately 38 years old, with values ranging from 31.0 to 44.6 years. The distribution is approximately symmetric, with one possible outlier of 31.0 years. The difference between the maximum median age and the minimum median age is 13.6 years.

Copyright 2025 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s). Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

3.19 **a.** Back-to-back (comparative) stem-and-leaf display (using only the first digit of each leaf)

Bills		Coins
9 7 7 6 5 5 5 4 3 3 3	0	2 3 3 5 6 8 9
6 4 2 2 1 1 0 0 0	1	0
7 5 4 0	2	2 2
4	3	2 5
	4	
	5	
3 0	6	

Variable

KEY: 6|3 = 633

b. The statement "Bills tend to be dirtier than coins" is generally consistent with the data. There is some overlap between the germ scores for the bills and coins, and the back-to-back stem-and-leaf plot shows that there are some very large germ scores for bills, which are not seen in the coins.

3.21 **a.**

0	9
1	6788999
2	00223334444445566678889
3	000112224568889
4	29
5	17

Stem: Tens
Leaf: Ones

b. The center is approximately 26 cents per gallon, and most states have a tax that is near the center value, with tax values ranging from 9 cents per gallon to 57 cents per gallon. The distribution is approximately symmetric. **c.** No. The only value that might be considered unusual is the 57 cents per gallon tax in Pennsylvania. Although not an obvious outlier, it is approximately 6 cents per gallon higher than the next lower gasoline tax. There are no other states that make such a big jump to the next higher tax. The lowest gasoline tax is 9 cents per gallon (Alaska) and the highest is 57 cents per gallon (Pennsylvania).

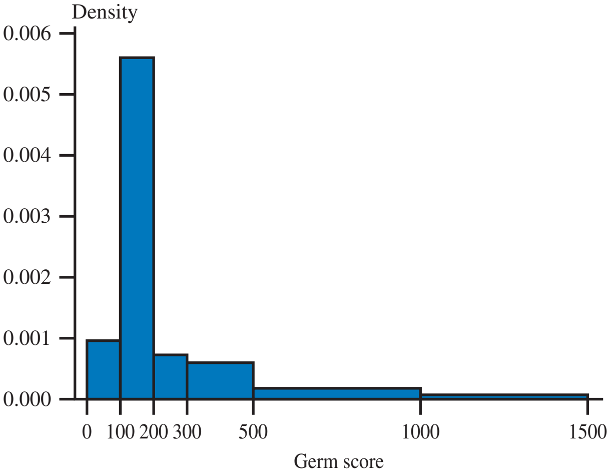
3.23

0f	44455555555
0s	6666666666666677777777
0*	8888888899999
1	0000

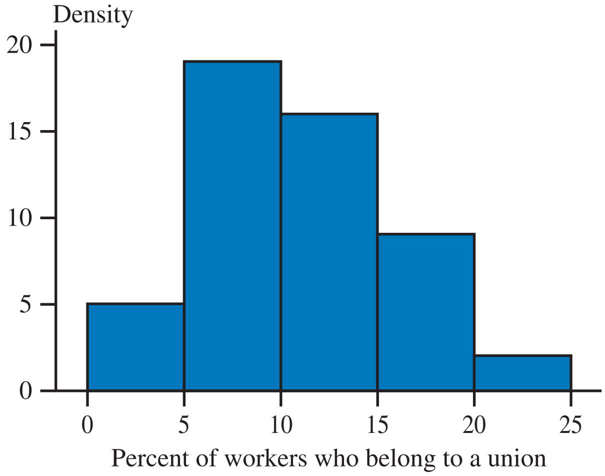
Stem: Tens
Leaf: Ones

The stem-and-leaf display shows that the distribution of high school dropout rates is roughly symmetrical. A typical dropout rate is 6–7%. The great majority of rates are between 4% and 10%, inclusive.

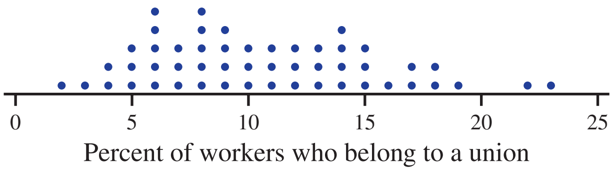
3.25



3.27 **a.**



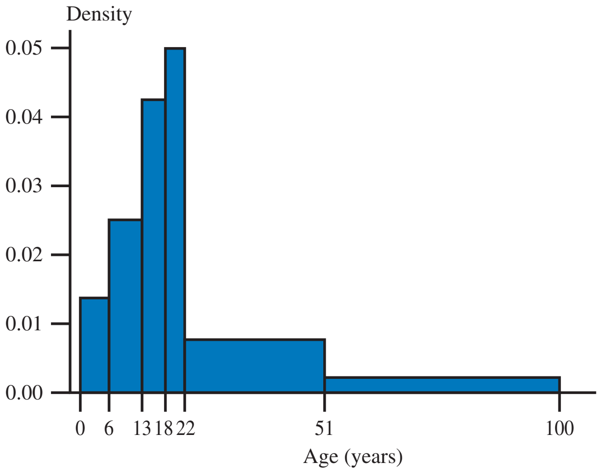
b.



A typical percentage of workers belonging to a union is around 10, with values ranging from 2.0 to 23.4. There are two states whose percentages stand out as being higher than those of the rest of the states. The distribution is positively skewed.

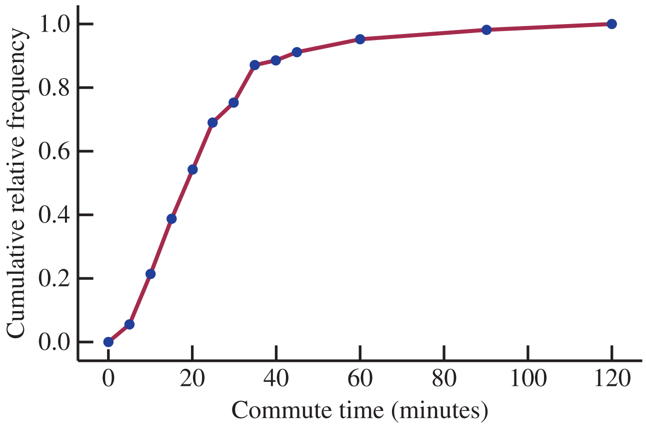
c. The dotplot is more informative as it shows where the data points actually lie. For example, in the histogram, we can tell that there are two observations in the 20 to 25 interval, but we don't see the actual values and miss the fact that these values are actually considerably higher than the other values in the data set.

3.29



Given that the eye is naturally drawn to large areas, the incorrect histogram exaggerates the proportion of clients in the wider age groups.

3.31 **a.**



b. i. 0.93 **ii.** $1 - 0.62 = 0.38$ **iii.** approximately 19 minutes

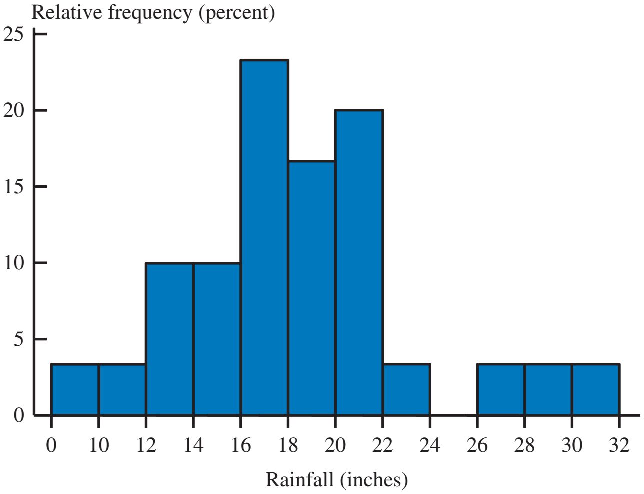
3.33 **a.** The distribution is likely to be negatively skewed. There will be a high density of scores close to 100, with no score greater than 100, and the density of scores tailing off to the left. **b.** So long as there is no cutoff at the lower end, the distribution is likely to be roughly symmetrical. For example, the greatest density of points might be around 65, with the density tailing off equally to the left and to the right of that value. **c.** In this case a bimodal distribution would be likely. There would be a clustering of points around, say, 65 for the students with less math knowledge and another clustering around, say, 95, for those who have had calculus.

3.35 **a.**

Rainfall (inches)	Frequency	Relative Frequency	Cumulative Relative Frequency
8 to < 10	1	0.0333	0.0333
10 to < 12	1	0.0333	0.0667
12 to < 14	3	0.1000	0.1667
14 to < 16	3	0.1000	0.2667
16 to < 18	7	0.2333	0.5000
18 to < 20	5	0.1667	0.6667
20 to < 22	6	0.2000	0.8667
22 to < 24	1	0.0333	0.9000
24 to < 26	0	0.0000	0.9000
26 to < 28	1	0.0333	0.9333
28 to < 30	1	0.0333	0.9667
30 to < 32	1	0.0333	1.0000

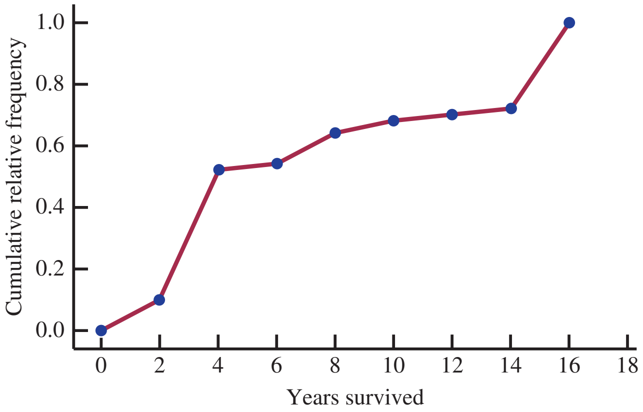
(continued)

b.



The histogram shows a distribution that is slightly positively skewed with a gap between 24 and 26.

3.37 **a.**



b. i. 0.53 **ii.** 0.62 **iii** $1 - 0.68 = 0.32$

3.39 Answers will vary. One possibility for each part is shown below.

a.

Class Interval	100 to <120	120 to <140	140 to <160	160 to <180	180 to <200
Frequency	5	10	40	10	5

Copyright 2025 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s). Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

b.

Class Interval	100 to <120	120 to <140	140 to <160	160 to <180	180 to <200
Frequency	20	10	4	25	11

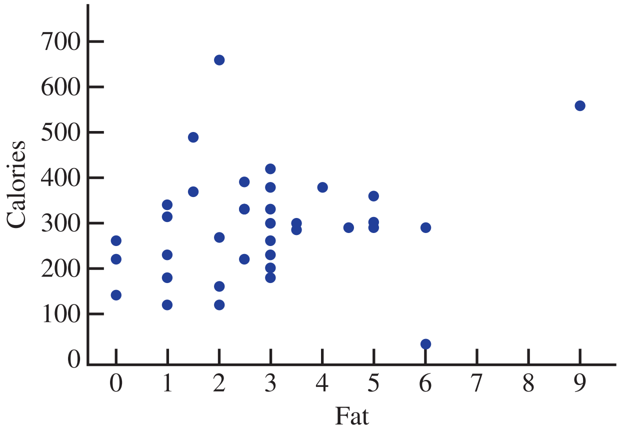
c.

Class Interval	100 to <120	120 to <140	140 to <160	160 to <180	180 to <200
Frequency	33	15	10	7	5

d.

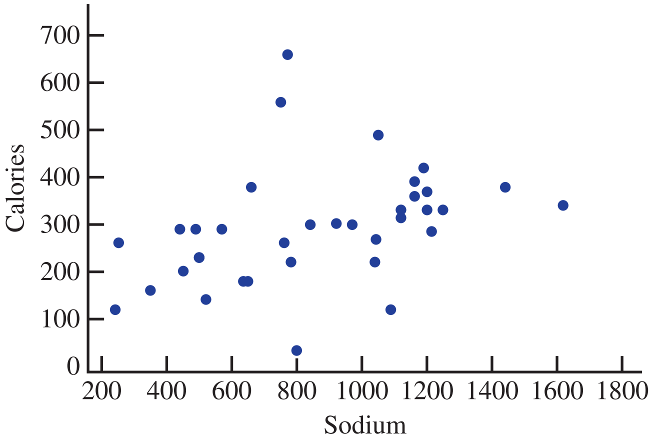
Class Interval	100 to <120	120 to <140	140 to <160	160 to <180	180 to <200
Frequency	5	7	10	15	33

3.41 **a.**



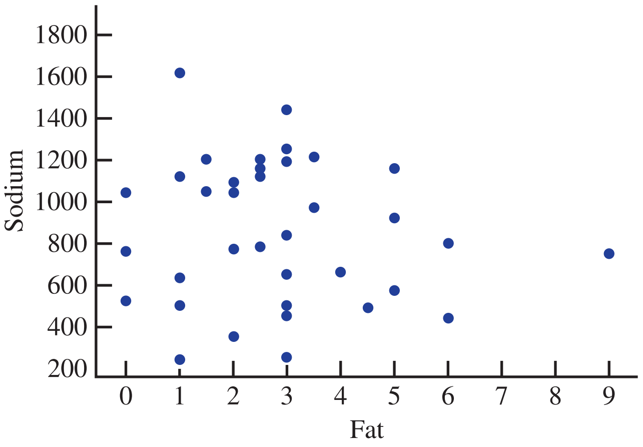
There does appear to be a relationship between fat and calories. As fat increases, calories also tends to increase, which is not surprising.

b.



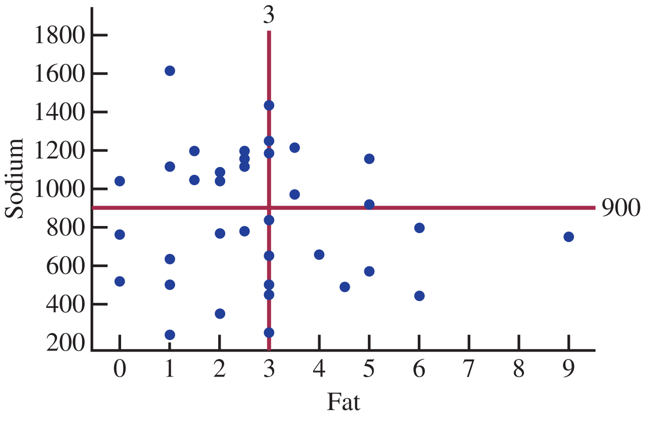
The relationship between sodium and calories is not strong, but it does appear that fast-food items with more sodium also tend to have more calories.

c.



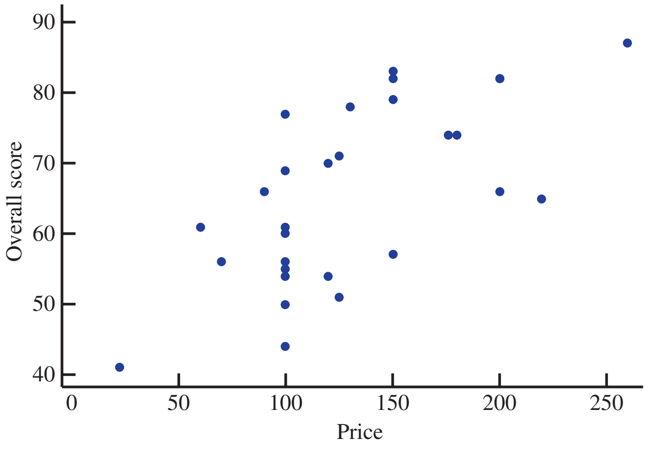
There does not appear to be a relationship between fat and sodium.

d.



The region in the lower left hand corner corresponds to the healthier fast-food choices. These are the choices that are lower in both sodium and fat.

3.43 **a.**



b. There appears to be a weak positive relationship between price and overall score. The scatterplot only weakly supports the statement.

3.45 The time series plot is consistent with the statement of having seen steady growth in recycling and composting because those trends are both increasing. However, the statement about the amounts by which landfills have generally declined might be somewhat misleading, at least when the conclusion is based on the graph shown, with the given scale. It might be more appropriate to state that the amounts of landfills have remained roughly constant, with a slight decrease in the given time period.

3.47 **a.** One pattern is that the number of complaints seems to be higher in the months when there is more snowfall. **b.** The authors might have chosen to include median snowfall data in the graph to make a connection between the number of complaints and some event that might be associated with the number of complaints. In this case, it is median snowfall data.

3.49 **a.** There are two observations that are possible unusual values, namely, 31 and 45 years old. The distribution is approximately symmetric, centered at about 38–39 years old, and has data values that vary between approximately 31 and 45 years. **b.** The graphical display shows the geographic distribution of median ages by state, and we see, for example, that states in the East tend to have higher median ages than those in the Southwest.

3.51 **a.** This solution considers the global temperature change. The spiraling motion to represent global temperature change allows the display to account for seasonal/monthly variation in temperatures. **b.** One revolution around the spiral represents 1 year. **c.** The colors in the animation indicate levels of severity of temperature change. As the global temperature change approaches a global average of 1.5 degrees Celsius, generally considered a threshold level for significant problems caused by that much of a temperature change, the colors get brighter. **d.** The animated graphic tells how, over time, the average global temperature has changed, including seasonal changes. **e.** The display does a good job in communicating its message. The display isn't overly complicated and is easy to follow and understand.

3.53 **a.** The variables represented are year, institution type (public 4-year, private nonprofit 4-year, or public 2-year), retention rate, and persistence rate. **b.** The colors represent the institution type (public 4-year, private nonprofit 4-year, or public 2-year). The colors make it easier to compare retention and persistence rates. For example, the light blue makes it easier to compare retention and persistence rates for private nonprofit 4-year institutions. **c.** Answers will vary. Over the 3-year period of 2017, 2018, and 2019, both retention and persistence rates remained roughly constant within each institution type. Note that the retention rate for public 4-year institutions increased slightly from 2018 to 2019, whereas private nonprofit 4-year and public 2-year institutions saw slight drops. Persistence rates tend to be higher than the retention rates for each institution type. **d.** The display does a good job in communicating its message. Although there are multiple variables represented, the display is not overly complex and concisely presents the data.

3.55 **a.** The variables represented in this display are geographic location and noise levels. We can learn where in the United States noise levels are higher and where they are lower. We can also see associations between urban and rural regions. **b.** Answers will vary. My geographic region (Eastern United States) has higher noise levels than much of the Western United States.

3.57 Answers will vary. There are many graphs that can be selected for analysis.

3.59 Answers will vary. Possible answers could include the countries that the recipients of Kiva's early loans were from compared to later loans, from what countries did the loans originate, what were the loans used for (agriculture, food, etc.), and how the number of loans changed over time.

3.61 Answers will vary. Statehood for Puerto Rico is more supported than statehood for Washington, DC as shown in the graph. Forty-three percent (43%) of respondents favor statehood for Puerto Rico, compared with only 36% favoring statehood for Washington, DC. Additionally, a greater percent of respondents oppose statehood for Washington, DC (37%) when compared with those who oppose statehood for Puerto Rico (31%).

3.63 **a.** Answers will vary. Over time, comparing 2002 to 2022, the smoking rates have decreased. In 2002, the highest smoking rate (roughly 35%) was in the 18–29 age group, and subsequent age groups show a steady decrease, with the lowest smoking rate (roughly 13%) in the 65+ age group. In 2022, the smoking rates for all age groups are lower than in 2002. However, the shape of the distribution is different. In 2022, the highest smoking rate (roughly 18%) is in the 50–64 age group, and the 30–49 age group has a smoking rate of approximately 17%. As with 2002, the lowest smoking rate is in the 65+ age group. **b.** Yes, the conclusion in the title of the article is accurate. One piece that is missing from the title is related to comparing the shapes of the distributions. Something that could be added in the title is an indication that smoking rates in 2002 steadily decrease as the age groups increase, whereas in 2022, the 30–49 and 50–64 age groups have the highest smoking rates.

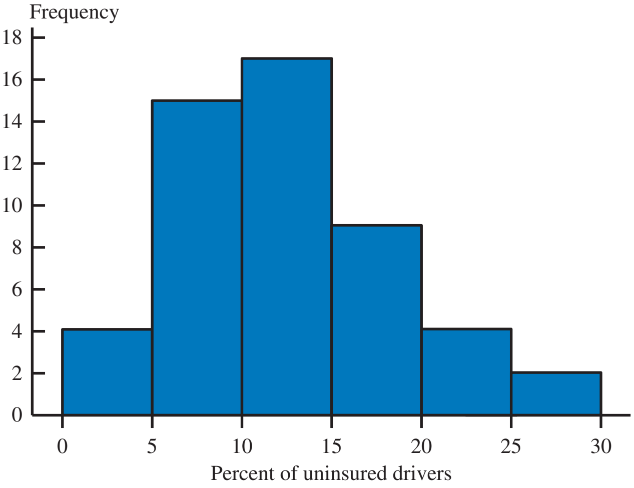
3.65 The first graph is a better representation of the incarceration rates. Although there is not a scale on the horizontal axis, the length of the bars are proportional to the incarceration rate. Of note are the bars for the United States and Iowa, which have much higher incarceration rates than the rates for the Founding NATO countries. In the second graph, the size of the bars for United States and Iowa are not nearly long enough given the horizontal scale.

3.67 Answers may vary. In comparing the percentages of Internet use among U.S. teens between 2014–2015 and 2022, it is clear that, based on this study, the Internet use has increased significantly. In 2014–2015, 80% of

Copyright 2025 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s). Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

respondents indicated that they use the Internet “Several times a day” or “Almost constantly.” In 2022, that percentage is 94%, a notable increase.

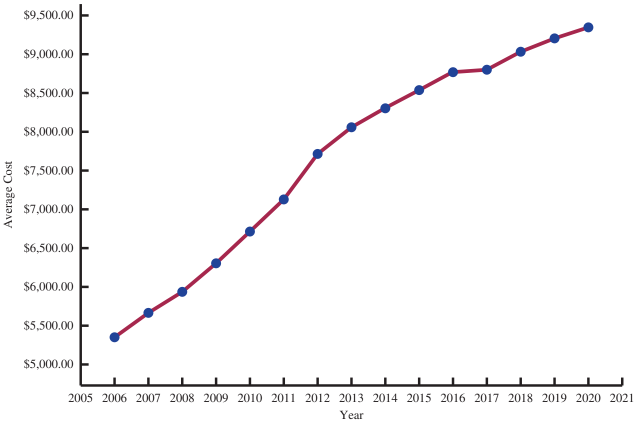
3.69



The distribution of the percent of uninsured drivers is positively skewed, centered in the 10–15% interval. The percent of uninsured drivers varies between a low of 3.1% and a high of 29.4%. Approximately 71% of states have 15% or fewer uninsured drivers.

3.71 **a.** High graft weight ratios are clearly associated with low body weights (and vice versa), and the relationship is not linear. (In fact, roughly speaking, there seems to be an inverse proportionality between the two variables, apart from a small increase in the graft weight ratios for increasing body weights among those recipients with the greater body weights. This is interesting in that an inverse proportionality between the variables would imply that the actual weights of transplanted livers are chosen independently of the recipients’ body weights.) **b.** A likely reason for the negative relationship is that the livers to be transplanted are probably chosen according to whatever happens to be available at the time. Therefore, lighter patients are likely to receive livers that are too large and heavier patients are likely to receive livers that are too small.

3.73



Over the time interval 2006 to 2020, the average cost per year for tuition and fees for four-year public universities in the United States has increased steadily. There is a strong approximately linear increase in average cost over time.

Cumulative Review Exercises 3

CR3.1 **a.** This is an experiment due to the random assignment of subjects to experimental conditions (the five different rooms). **b.** This is an observational study because there was no random assignment of subjects to experimental conditions; the researchers merely recorded what they observed about colon cancer rates for the nurses who consumed more than two servings of sugar-sweetened beverages per week versus those who consumed less than one sugar-sweetened beverage per week. **c.** This is an observational study because there was no random assignment of subjects to experimental conditions; the researchers merely recorded the responses of the survey participants. **d.** This is an experiment because of the random assignment of study participants (the adults with back pain) to experimental conditions (the four different treatments).

CR3.3 Very often those who choose to respond generally have a different opinion on the subject of the study from those who do not respond. (In particular, those who respond often have strong feelings against the status quo.) This can lead to results that are not representative of the population that is being studied.

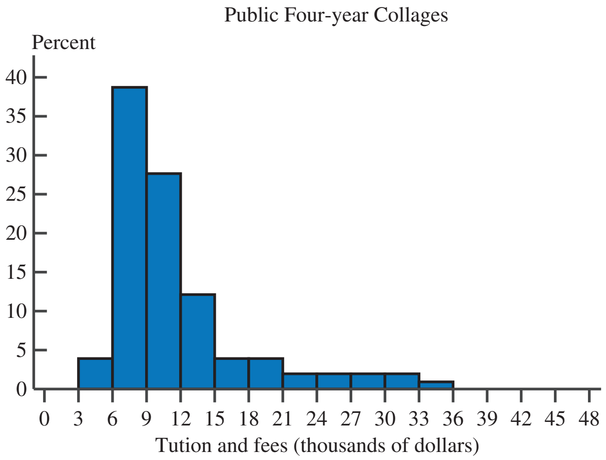
CR3.5 **a.** Double-blind refers to neither the study participants nor those measuring the response or providing the treatment know what treatment was received. The purpose of blinding is to not let personal beliefs influence the measurements or inadvertently, perhaps through facial expressions, offer encouragement to some patients and not to others by knowing what the treatment is. For the study participants, blinding helps reduce the placebo effect. **b.** It might be reasonable to conclude that the doctors’ belief about the effectiveness of the pain reliever is what caused the observed differences in patient pain. Perhaps the doctors provided facial cues that led the patient to believe the treatment was a pain reliever. **c.** It isn’t reasonable to generalize the results of this study to all doctors and patients because there was no random selection from the population of doctors and patients.

CR3.7 **a.** Gender, Brand of motorcycle, Telephone area code **b.** Number of previous motorcycles owned by purchaser **c.** Bar chart **d.** Dotplot

CR3.9 **a.** I would use rate per 10,000 flights data to rank airlines based on flights delayed on the tarmac for at least 3 hours. The reason for this is because the airlines are of different sizes, and have different total numbers of flights. In order to compare groups of different sizes, the airlines must be put on the same scale, namely, rate per 10,000 flights. **b.** Of the 17 airlines in the study, 13 had fewer than 2 delays per 10,000 flights, and there are 2 airlines that had between 2.5 and 3 delays per 10,000 flights.

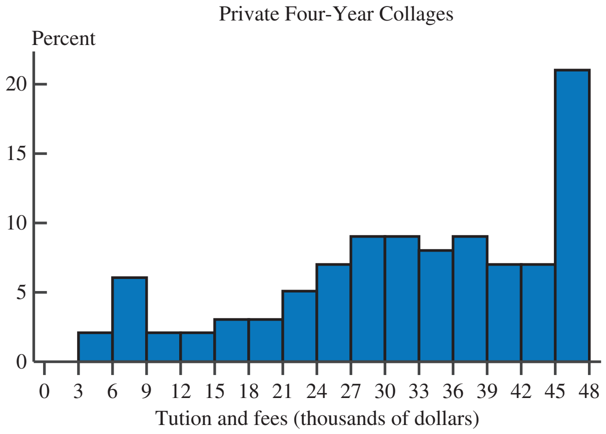
Two airlines, however, had much greater rates, namely, ExpressJet and Continental Airlines. These two airlines had the highest numbers of delays per 10,000 flights.

CR3.11 **a.**



The relative frequency histogram shows that the distribution of tuition and fees for public four-year colleges is positively skewed, with the most frequently occurring tuition and fee amounts between $6,000 and $9,000. The center of the distribution is in the range $9,000 to $12,000. The tuition and fees values range between a low of between $3,000 and $6000, to a high of somewhere between $33,000 and $36,000. There are no tuition and fee values above $36,000.

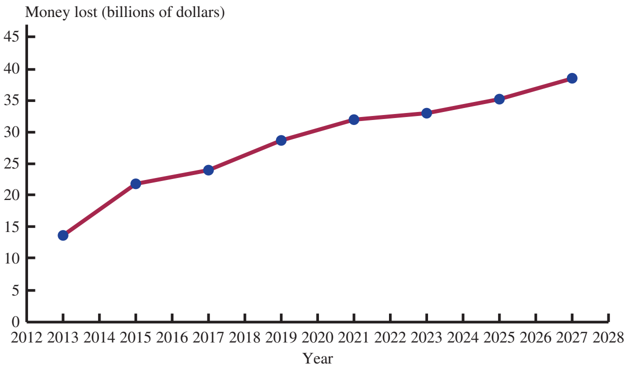
b.



The relative frequency bar chart shows that the distribution of tuition and fees for private nonprofit four-year colleges is negatively skewed, with the most frequently occurring tuition and fee amounts between $45,000 and $48,000. The center of the distribution is between $33,000 and $36,000. The tuition and fees values range between a low of between $3,000 and $6,000 to a high of between $45,000 and $48,000. There are no tuition and fee values below $3,000 or above $48,000.

c. The differences in the distributions of tuition and fees of public and private four-year colleges are in shape, center, and spread. The shapes are both skewed, with the public college tuition and fees distribution being positively skewed, and the private college tuition and fees distribution being negatively skewed. The centers are quite different, with the center of the private college tuition and fees higher than that for public colleges. Finally, spread of the tuition and fee amounts for private colleges is greater than that for public colleges, with the private colleges having higher tuitions than the public colleges.

CR3.13



The amount of money lost to credit card fraud has increased over time at a nearly steady rate. The time series plot does not show any decreases over time, and the slopes of the line segments from one point to the next are nearly the same.

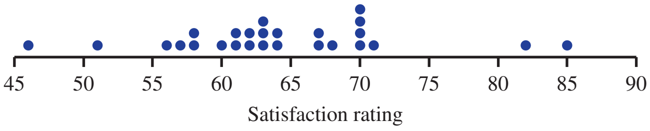
CR3.15 The time series plot is consistent with the statement of having seen steady growth in recycling and composting because those trends are both increasing. However, the statement about amounts landfilled have generally declined might be somewhat misleading, at least when the conclusion is based on the graph shown, with the given scale. Rather, it might be more appropriate to state that the amounts landfilled have remained roughly constant, with a slight decrease in the given time period.

Chapter 4

4.1 **a.** $\bar{x} = \$148.77$; median $= \$142.33$. **b.** The mean is larger than the median since the distribution of these six values is positively skewed. The largest value is greatly separated from the remaining five values. **c.** The median is better as a description of a typical value since it is not influenced by the extreme value.

4.3 The mean caffeine concentration for the brands of coffee listed $= 125.417$ mg/cup. Therefore, the mean caffeine concentration of the coffee brands in mg/ounce is 15.677. This is significantly greater than the previous mean caffeine concentration of the energy drinks.

4.5 **a.**



The distribution has unusual observations at both the low and high end of the Satisfaction Rating scale that seem somewhat more removed from the majority of the data near the center. The distribution is approximately symmetric.

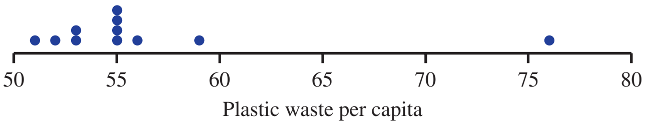
b. $\bar{x} = (85 + \cdots + 46)/26 = 64.19$

Median = average of two middle values = 63

c. The mean and median are similar because the distribution is approximately symmetric.

4.7 **a.** There is an even number of observations ($n = 10$), so the median is the average of the two middle values in the ordered data set, which is 177,746.5. This is the value that divides the ordered data set into two halves. This tells us that half of the values in the data set had average weekly circulations of less than 177,746.5, and the other half had average weekly circulations of more than 177,746.5. **b.** The median is preferable to the mean for describing the center for this data set because the distribution is positively skewed and contains outliers. **c.** It is not reasonable to generalize from this sample to the population of daily newspapers in the United States because these newspapers were not randomly selected. Rather, they are the top 10 newspapers in weekday circulation.

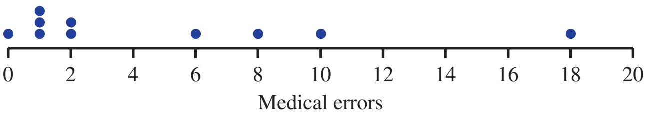
4.9 **a.** The distribution of plastic waste generation per capita values is positively skewed (see dotplot below), so the median should be used to describe the center for this data set.



b. It is not reasonable to draw conclusions about per-capita single-use plastic waste generation for all countries in the world based on the given sample because this is not a random sample of countries. Rather, this sample contains data for the 11 countries with the highest single-use waste generation, which are likely not representative of all countries in the world.

4.11 **a.** The median estimated total cost per year is \$9,119. **b.** The median is preferable to the mean because the distribution of estimated total cost is positively skewed (there are possible outliers or at least some unusually large values at the higher end of the distribution). **c.** The mean is likely greater than the median because the distribution is positively skewed.

4.13 **a.** The dotplot is positively skewed. The mean and median will be different from each other, with the mean being greater than the median.



b. $\bar{x} = 4.9$ errors, median $= \dfrac{2+2}{2} = 2$ errors **c.** The largest value in the data set could be increased by any amount and not affect the sample median because the position of the middle value will not change if the largest value is increased. The largest value could be decreased to 2 errors without changing the value of the median.

4.15 The median is 680 hours.

4.17 **a.** $\bar{x} = 52.111$. Variance = 279.111. $s = 16.707$. **b.** The addition of the very expensive cheese would increase both the mean and the standard deviation.

4.19 The median is 1012.5 milliseconds. The lower quartile is 982 milliseconds, and the upper quartile is 1052 milliseconds. The interquartile range is $iqr = 1052 - 982 = 70$ milliseconds. Half of the reaction times are less than 1012.5 milliseconds, and half of the reaction times are greater than 1012.5 milliseconds. The middle 50% of the data values have a range of 70 milliseconds.

4.21 The mean is $\bar{x} = 3510$ grams. This is a typical or representative value for the body mass of the 10 female chinstrap penguins. The standard deviation is $s = 211.87$ grams, which represents how much, on average, the values in the data set spread out, or deviate, from the mean.

4.23 **a.** Variance = 22,735.7.
Standard deviation $= \sqrt{22735.7} = 150.78$
b. The large value of the standard deviation tells us that there is considerable variation between prices of these highly rated smart phones.

4.25 Variance = 4922.0
$s = \sqrt{4922.0} = 70.2$ milliseconds

4.27 **a.** Lower quartile = 0. Upper quartile = 195. Interquartile range = 195. **b.** The lower quartile equals the minimum value for this data set because there is a large number of equal values (zero in this case) at the lower end of the distribution. This is unusual, and therefore, generally speaking, the lower quartile is not equal to the minimum value.

4.29 **a.** The standard deviation of percent return is a reasonable measure of unpredictability because it measures how much, on average, individual returns deviate from the mean return of the fund. A smaller standard deviation indicates smaller deviations (on average) from the mean return, and therefore less risk. **b.** A fund with a small standard deviation can still lose money if the average percent return is small relative to the standard deviation. Recall that the standard deviation is a typical deviation from the mean (either above or below the mean), and if the mean is smaller than the standard deviation, the return could be negative, resulting in the fund losing money.

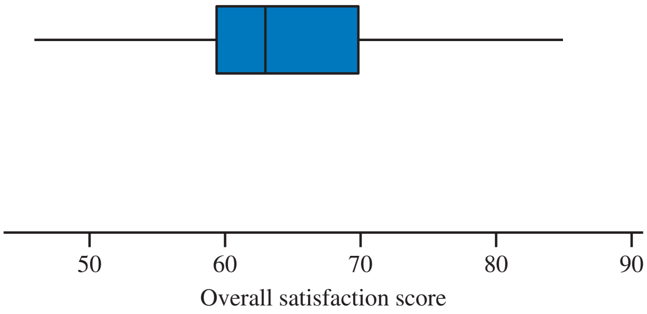
4.31 **a.** i **b.** iii **c.** iv

4.33 **a.**

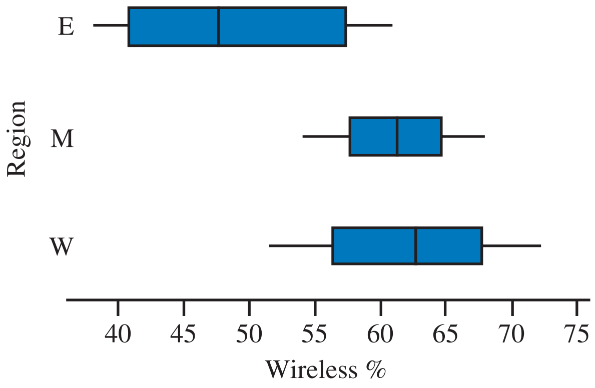
	Mean	Standard Deviation	Coefficient of Variation
Sample 1	7.81	0.398	5.102
Sample 2	49.68	1.739	3.500

b. The values of the coefficient of variation are given in the table in Part (a). The fact that the coefficient of variation is smaller for Sample 2 than for Sample 1 is not surprising since, relative to the actual amount placed in the containers, it is easier to be more accurate when larger amounts are being placed in the containers.

4.35 **a.** lower quartile = 60, upper quartile = 70. *iqr* = 70 − 60 = 10. **b.** Any values greater than upper quartile + 1.5(*iqr*) or less than lower quartile − 1.5(*iqr*) are considered outliers. For this data set, values are outliers if they are greater than 70 + 1.5(10) = 85 or less than 60 − 1.5(10) = 45. There are no values greater than 85 or less than 45, so there are no outliers in this data set. **c.** The boxplot shows no outliers. The distribution of overall satisfaction scores is slightly positively skewed.



4.37 **a.**



b. The Eastern region has the smallest median (47.7%), and the Middle and Western regions have medians that are much closer to each other (61.2% and 62.7%, respectively). The distribution of wireless percent for the West is negatively skewed, the distribution of wireless percent for the East is positively skewed, and that for the Middle is approximately symmetric. The Eastern region has the most variability in wireless percent, and the Middle region has the least variability.

4.39 **a.** Median = average of 9th and 10th values = (10 + 10)/2 = **10**

Lower quartile = 5th value = **6**

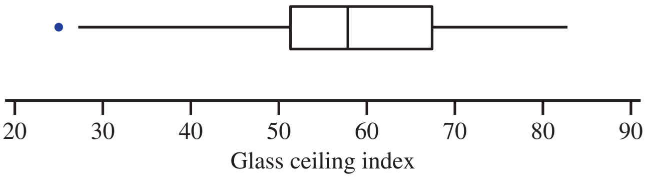
Upper quartile = 14th value = **13**

Interquartile range = 13 − 6 = **7**

b. There are no outliers.

4.41 **a.** Lower quartile: 51.3; Upper quartile: 67.45; Interquartile range = 67.45 − 51.3 = 16.15. South Korea's observation (25.0) is an outlier.

b.



c. The Nordic countries are located in the whisker at the right side of the boxplot.

4.43 **a.** Roughly 68% of speeds would have been between those two values. **b.** Roughly 16% of speeds would exceed 57 mph.

4.45 Since the mean of the distribution is 32 minutes and the standard deviation is 15 minutes, 0 is just over 2 standard deviation below the mean. Therefore, if the distribution were approximately normal, then just under (100 − 95)/2 = 2.5% of travel times would be less than 0, which is clearly not the case. Thus the distribution cannot be well approximated by a normal curve.

4.47 **a.** At least 75% of observations must lie between those two values. **b.** The required interval is (2.90, 70.94). **c.** The distribution cannot be approximately normal.

4.49 For the first test, $z = 1.5$; for the second test, $z = 1.875$. Since the student's z score in the second test is higher, the student did better relative to the other test takers in the second test.

4.51 The 20th percentile of $5,000 indicates that 20% of students have $5,000 or less of student debt. The 36th percentile indicates that 36% of students have $10,000 or less of student debt. The 63rd percentile indicates that 63% of students have $25,000 or less of student debt. The 82nd percentile indicates that 82% of students have $50,000 or less of student debt. The 90th percentile indicates that 90% of students have $75,000 or less of student debt.

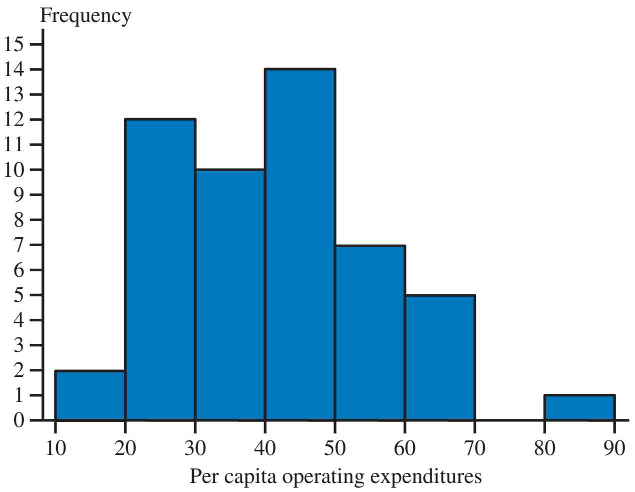
4.53 The best conclusion we can reach is that at most 16% of weight readings will be between 49.75 and 50.25.

4.55 The value of the standard deviation tells us that a typical deviation of *the number of answers changed from right to wrong* from the mean of this variable is 1.0. However, 0 is only 0.9 below the mean and negative values are not possible, and so for a typical deviation to be 1.0 there must be some values more than 1.0 *above* the mean, that is, values above 1.9. This suggests that the distribution is positively skewed.

The value 3 is the lowest whole number value more than 3 standard deviations above the mean. Therefore, using Chebyshev's Rule, we can conclude that at most $1/3^2 = 1/9$ of students; that is, at most 72/9 = 8 students changed at least three answers from correct to incorrect.

Copyright 2025 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s). Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

4.57 **a.**



b. i. The 50th percentile is between per capita operating expenditures of 40 and 50. **ii.** The 70th percentile is also between per capita operating expenditures of 40 and 50. **iii.** The 10th percentile is between per capita operating expenditures of 20 and 30. **iv.** The 90th percentile is between per capita operating expenditures of 60 and 70. **v.** The 40th percentile is between per capita operating expenditures of 30 and 40.

4.59 **a.** The minimum value and the lower quartile were both 1. **b.** More than half of the data values were equal to the minimum value. **c.** Between 25% and 50% of patients had unacceptable times to defibrillation.
d. (Upper quartile) + 3(iqr) = 9. Since 7 is less than 9, 7 must be a mild outlier.

4.61 **a.** $\bar{x} = 299$. The five deviations are: 46, -7, 35, -23, -51 **b.** The sum of the rounded deviations is 0.
c. Variance = 1630
$s = \sqrt{1630} = 40.37$

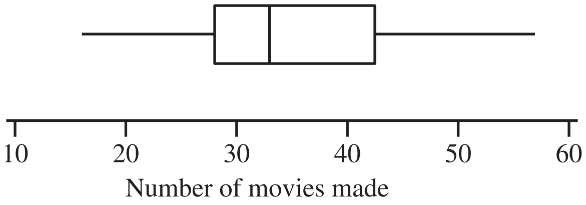
4.63 **a.** This is a correct interpretation of the median. **b.** Here the word "range" is being used to describe the interval from the minimum value to the maximum value. The statement claims that the median is defined as the midpoint of this interval, which is not true. **c.** If there is no home below $300,000, then certainly the median will be greater than $300,000 (unless more than half of the homes cost exactly $300,000).

4.65 The new mean is $\bar{x} = 38.364$. The new values and their deviations from the mean are shown in the table below.

Value	Deviation
52	13.636
13	−25.364
17	−21.364
46	7.636
42	3.636
24	−14.364
32	−6.364
30	−8.364
58	19.636
35	−3.364

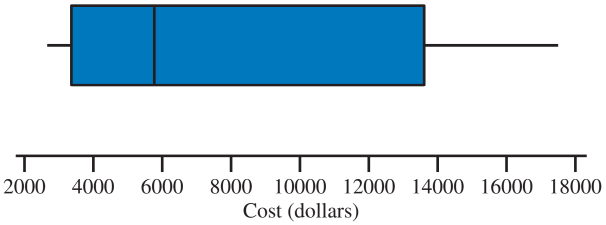
The value of s^2 for the new values is the same as for the old values.

4.67



The distribution of number of movies made is slightly positively skewed. This tells us that there is less variability in the lower 50% of number of movies made than in the upper 50%.

4.69 **a.** Sample mean: $\bar{x} = 7687.73$; Sample standard deviation: $s = 5065.56$ **b.** $Q_3 = 13610$; $Q_1 = 3350$; $IQR = 13610 - 3350 = 10260$ **c.** Any values greater than $Q3 + 1.5(IQR) = 13610 + 1.5(10260) = 29000$ are considered outliers. Since the cost for the United States is 13610, it is not considered an outlier.
d.



Chapter 5

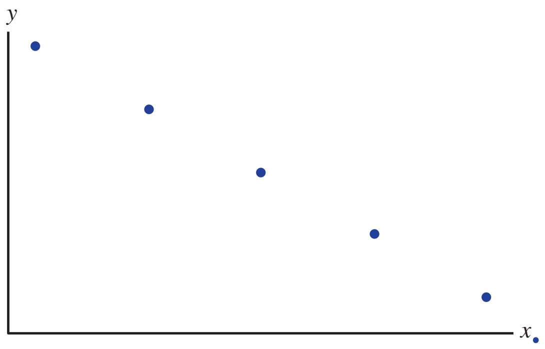
5.1

Scatterplot 1	**(i)** Yes	**(ii)** Yes	**(iii)** Positive
Scatterplot 2	**(i)** Yes	**(ii)** Yes	**(iii)** Negative
Scatterplot 3	**(i)** Yes	**(ii)** No	**(iii)** –
Scatterplot 4	**(i)** Yes	**(ii)** Yes	**(iii)** Negative

5.3 **a.** Positive. Older children tend to be taller than younger children. **b.** Close to zero. There is no reason to believe that there is an association between height and score on a college entrance exam. **c.** Positive. People with large feet tend to be taller than people with small feet.

5.5 No. A correlation coefficient of 0 implies that there is no *linear* relationship between the variables. There could still be a nonlinear relationship.

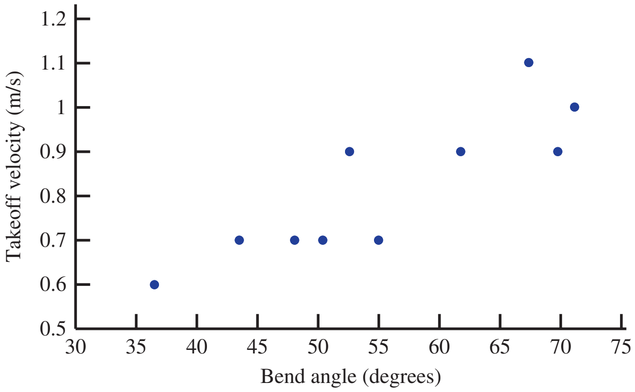
5.7 Scatterplot for which $r = -1$:



5.9 **a.** $r = 0.204$. There is a weak positive linear relationship between cost per serving and fiber per serving. **b.** $r = 0.241$. This correlation coefficient is slightly greater than the correlation coefficient for the per serving data.

5.11 **a.** The value of the correlation coefficient is negative, which suggests that students who use a cell phone for more hours per day tend to have lower GPAs. **b.** The relationship between texting and GPA has the same direction as the direction for cell phone use and GPA, so the correlations must have the same sign. The relationship between texting and GPA is not as strong as the relationship between cell phone use and GPA, so the correlation coefficient must be closer to zero. Therefore, $r = -0.10$ is the only option that satisfies both of the criteria. **c.** Since it is reasonable to believe that texts sent would be approximately equal to texts received, there would be a positive association between these items. In addition, given that the two texting items are nearly perfectly correlated, the correlation coefficient must be close to $+1$ or -1. Therefore, the correlation coefficient must be close to $+1$.

5.13 **a.**



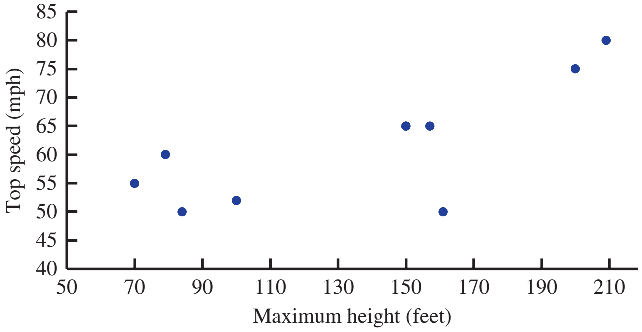
There is a strong, positive, linear relationship between bend angle and takeoff velocity. **b.** The correlation of $r = 0.867$ indicates that there is a strong, linear relationship with a positive association between bend angle and takeoff velocity. **c.** The correlation of 0.867 indicates a strong linear relationship between bend angle and takeoff velocity, the positive correlation indicates a positive association.

5.15 No, because "artist" is a categorical variable.

5.17 Scatterplot 1 seems to show a linear relationship between x and y, while Scatterplot 2 shows a curved relationship between the two variables. It makes sense to use the least-squares regression line to summarize the relationship between x and y for the data set in Scatterplot 1, but not for the data set in Scatterplot 2.

5.19 **a.** $\hat{y} = 51.305 + 0.1633x$, where $\hat{y}$ is the predicted percentage of alumni who would strongly agree and x is the ranking. **b.** 59.47 **c.** Since the ranking of 10 is outside this range, the least-squares regression line should not be used to predict the percentage of alumni who would strongly agree because there is no guarantee that the linear pattern will continue outside this range.

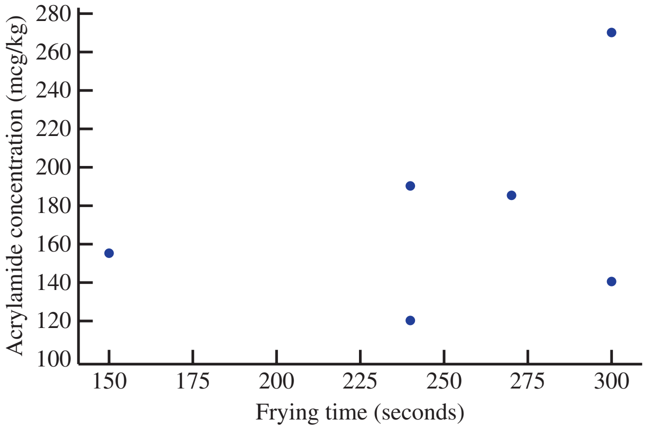
5.21 **a.** Top speed is the response variable (y), and maximum height is the predictor variable (x). **b.**



There is a moderately strong, positive linear relationship between top speed and maximum height. **c.** $\hat{y} = 40.213 + 0.158x$, where $\hat{y}$ is the predicted top speed and x is the maximum height. **d.** For a 1-foot increase in maximum height, the predicted top speed increases by 0.158 mph. **e.** It does not make sense to interpret the intercept of the least-squares regression line. A maximum height of zero feet is far outside the range of maximum height values that was used to develop the least-squares regression line. **f.** The predicted top speed for a roller coaster with a maximum height of 150 feet is $\hat{y} = 40.213 + 0.158(150) = 63.913$ mph. **g.** It is not reasonable to use the least-squares regression line to predict the top speed of a roller coaster with a maximum height of 300 feet. Three hundred (300) feet is outside the range of maximum height values that were used to create the least-squares regression line.

5.23 **a.** Negative. As the patient-to-nurse ratio increases we would expect nurses' stress levels to increase and therefore their job satisfaction to decrease. **b.** Negative. As the patient-to-nurse ratio increases we would expect patient satisfaction to decrease. **c.** Negative. As the patient-to-nurse ratio increases we would expect quality of care to decrease.

5.25 **a.** The response variable is the acrylamide concentration, and the predictor variable is the frying time. **b.**



Copyright 2025 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s). Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

There is a weak, positive association ($r = 0.379$) between frying time and acrylamide concentration.

5.27 **a.** There is a strong negative relationship between mean call-to-shock time and survival rate. The relationship is close to linear, particularly if the point with the highest x value is disregarded. If that point is included, then there is the suggestion of a curve. **b.** $\hat{y} = 101.32847 - 9.29562x$

5.29 The slope would be 0.031, since the slope of the least-squares line is the increase in the average y value for each increase of one unit in x. Here the average electricity consumption increases by 0.031 kWH for each increase of 1 degree Celsius.

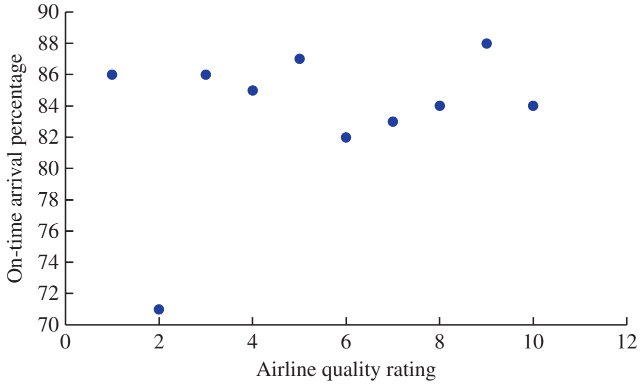
5.31 We do not know that the same linear relationship will apply for x values outside this range.

5.33 $b = r(s_y/s_x)$, where s_y and s_x are the standard deviations of the y values and the x values, respectively. Since standard deviations are always positive, b and r must have the same sign.

5.35 **a.** The equation of the least squares regression line is $\hat{y} = -9.071 + 1.571x$, where x = years of schooling and y = median hourly wage gain. When $x = 15$, $\hat{y} = 14.5$. The least squares regression line predicts a median hourly wage gain of 14.5 percent for the 15th year of schooling. **b.** The actual wage gain percent is very close to the value predicted in Part (a).

5.37 **b.** The scatterplot for girls shows stronger evidence of a curved relationship. **c.** $\hat{y} = 479.997 + 12.525x$ **e.** The decision to use a curve is supported by the curved pattern in the residual plot.

5.39 The pattern in the scatterplot looks somewhat linear. There is an unusual value (Allegiant Airlines at (2, 71)). Excluding the unusual value, the slope of the least-squares regression line is nearly zero.



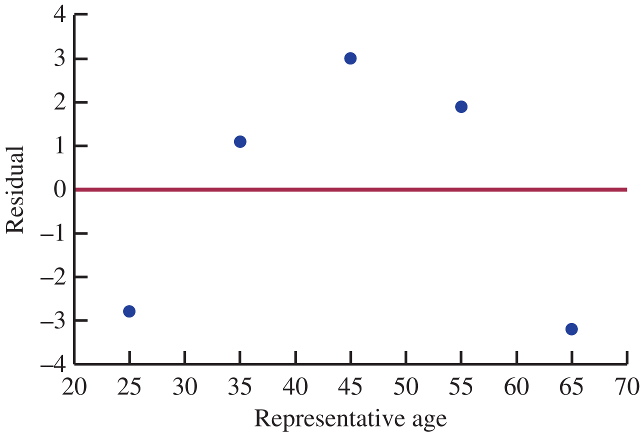
b. The regression equation is $\hat{y} = 80.933 + 0.4848x$, where $\hat{y}$ is the predicted on-time arrival percentage and x is the airline quality rating.
c. The residuals are

2.1278	−10.903	−1.3266	3.6126	−1.781	2.7038
−1.8418	4.5822	3.643	−0.8114		

5.41 **a.** The observation (150,155) is potentially influential because that point has an x value that is far away from the rest of the data set. **b.** $\hat{y} = -44 + 0.83(270) = 180.1$; This prediction is smaller than that made in the previous exercise.

5.43 **a.** The least-squares regression line equation is $\hat{y} = 37.05 - 0.09x$. **b.** The residuals are shown in the right-most column of the table.

Representative Age (x)	Average Stress Level (y)	Predicted Average Stress Level ($\hat{y}$)	Residual ($y - \hat{y}$)
25	32	34.8	−2.8
35	35	33.9	1.1
45	36	33	3
55	34	32.1	1.9
65	28	31.2	−3.2

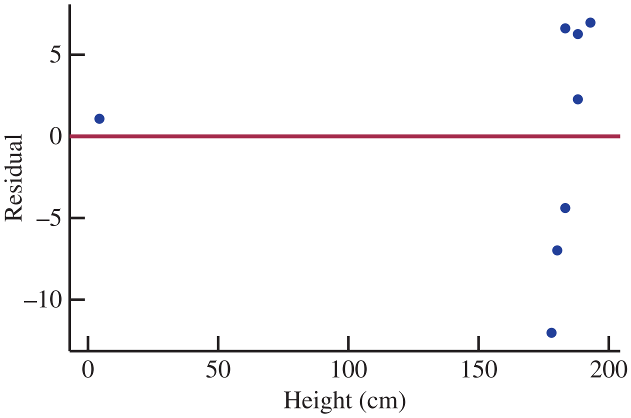


c. It is not recommended to use the least-squares regression line to describe the relationship between stress level and age. A curve is a better choice to describe the relationship. This is based on the obvious curved pattern in the residual plot.

5.45 **a.**

Height (x)	Weight (y)	Predicted Weight ($\hat{y}$)	Residuals ($y-\hat{y}$)
188	95	88.66	6.34
4	4	2.916	1.084
188	91	88.66	2.34
178	72	84	−12
183	93	86.33	6.67
180	78	84.932	−6.932

b.



The residual plot has one observation (the Lego Batman, at 4 cm and 4 kg) that is noticeably different from the others. There is a strong pattern in the residual plot. **c.** This point is influential in determining both the slope and the intercept. That point is far removed from the other points in both the x and y directions. In fact, by removing the Lego Batman point, the equation of the least-squares regression line changes to $\hat{y} = -220.43 + 1.6643x$.
d. No. The least-squares line in Part (a) was calculated using the data for men, and we have no reason to believe that this line will also apply to women.

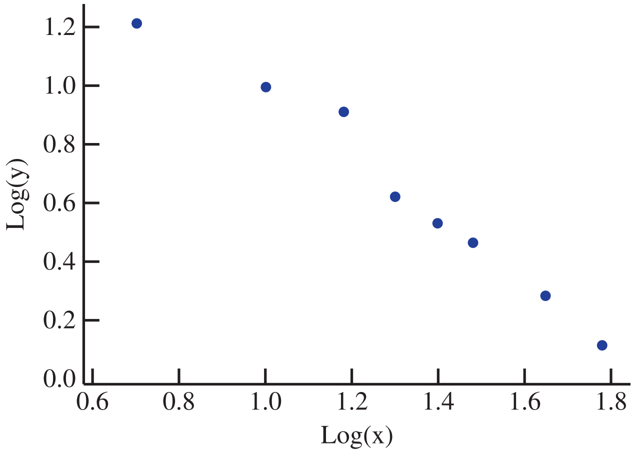
5.47 **a.** 0.154 **b.** No, since the r^2 value for y = first-year-college GPA and x = SAT II score was 0.16, which is not large. Only 16% of the variation in first-year-college GPA could be attributed to the approximate linear relationship between SAT II score and first-year-college GPA.

5.49 **b.** $\hat{y} = 85.334 - 0.0000259x$. $r^2 = 0.016$. **c.** The line will not give accurate predictions. **d.** Deleting the point (620,231, 67), the equation of the least-squares line is now $\hat{y} = 83.402 + 0.0000387x$. Removal of the point does greatly affect the equation of the line.

5.51 $r^2 = 0.951$. 95.1% of the variation in hardness is attributable to the approximate linear relationship between time elapsed and hardness.

5.53 **a.** $r = 0$, $\hat{y} = \bar{y}$. **b.** For values of r close to 1 or -1, s_e will be much smaller than s_y.

5.55 **a.**

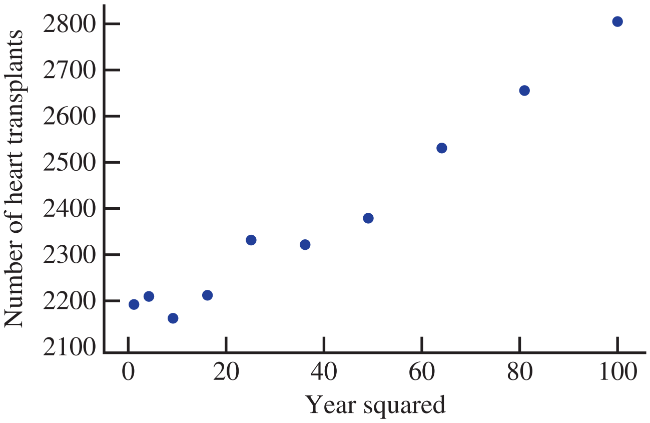


The relationship between log(x) and log(y) seems to be linear. **b.** The value of r^2 is large (0.973) and the value of s_e is relatively small (0.0657), indicating that a line is a good fit for the data. **c.** 2.477%

5.57 Answers will vary. Scatterplot (b) shows a curve that tends to follow the pattern of points in the scatterplot.

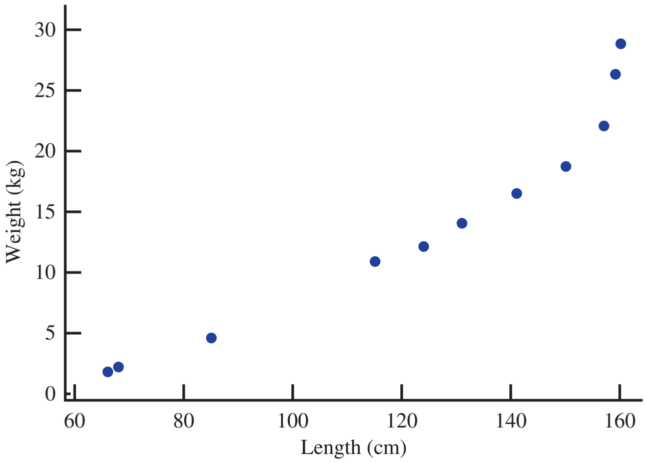
5.59 **a.** The relationship appears to be nonlinear.
b. There is an obvious pattern in the residual plot.
c. The second transformation (ln(x) vs. y) seems to be somewhat more successful in straightening the data.
d. $\hat{y} = 62.42 + 43.81x$, where $\hat{y}$ is the predicted success %, and x is the natural logarithm of the energy of shock.
e. 86.94%, 52.64%

5.61 **a.** One transformation to straighten the plot is x = year^2 and y = number of heart transplants.



b. The equation of the least-squares regression line is $\hat{y} = 2139.87 + 6.2321x^*$, where $\hat{y}$ is the predicted number of heart transplants, and x^* is the year (since 2006) squared. The predicted number of heart transplants in 2016 (the eleventh year since 2006, or $x^* = 11^2 = 121$) is $\hat{y} = 2139.87 + 6.2321(121) = 2893.9541$. It is predicted that there will be approximately 2894 heart transplants in 2016. **c.** We have to be confident that the pattern observed between 2006 and 2015 will continue up to 2016. This is reasonable so long as circumstances remain basically the same. To expect the same pattern to continue to 2026, however, might be unreasonable.

5.63

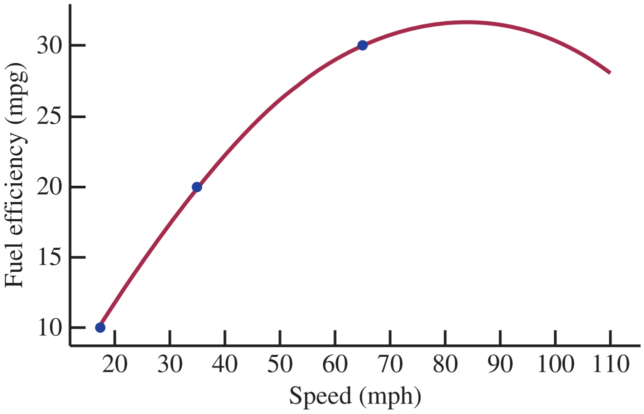


Copyright 2025 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s). Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

The scatterplot shows a nonlinear relationship. A transformation that makes the plot much closer to linear is $x' = \ln(\text{Length})$ and $y' = \ln(\text{Weight})$.

5.65 Consider the scatterplot for the combined group. Between $x = 0$ and $x = 1$ the points form an approximately linear pattern with a slope of 6.93, while between $x = 1$ and $x = 10$ the points form an approximately linear pattern with a slope of just over 3. Clearly this pattern could not be modeled with a single straight line; a curve would be more appropriate.

5.67

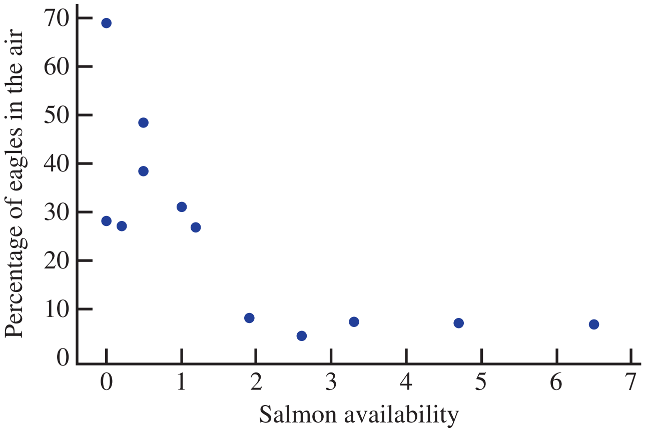


As illustrated in the graph above, it is quite possible that the relationship between speed and fuel efficiency is modeled by a curve (in particular, by a quadratic curve), and that for greater speeds the fuel efficiency is negatively related to the speed.

5.69 **a.** $r = -0.10$; there is a weak, negative linear relationship. Because the relationship is negative, larger arch heights tend to be paired with smaller average hopping heights. **b.** The correlation coefficients support the conclusion since they are all fairly close to 0.

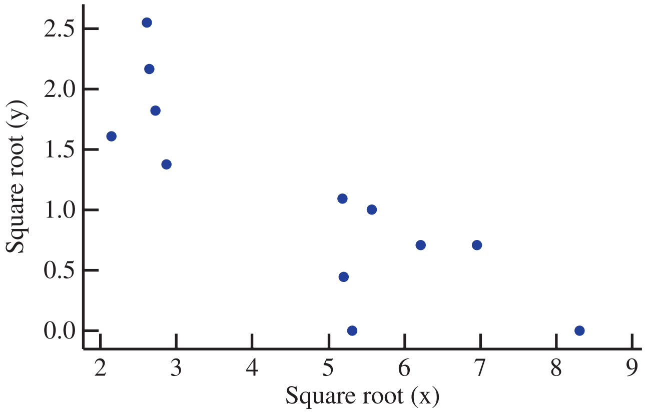
5.71 **a.** Positive, because the cost of produce is determined by its weight, so a heavier apple would cost more. **b.** Close to zero, because there is no reason to believe that the height of a person would be at all associated with the number of pets that a person has. **c.** Positive, because, in general, the more time spent studying for an exam, the higher the score will be.

5.73 **a.**



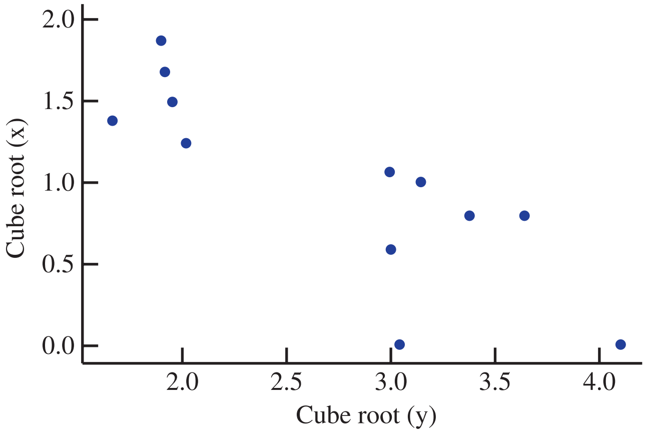
There seems to be a curved relationship between percentage of eagles in the air and salmon availability.

c.

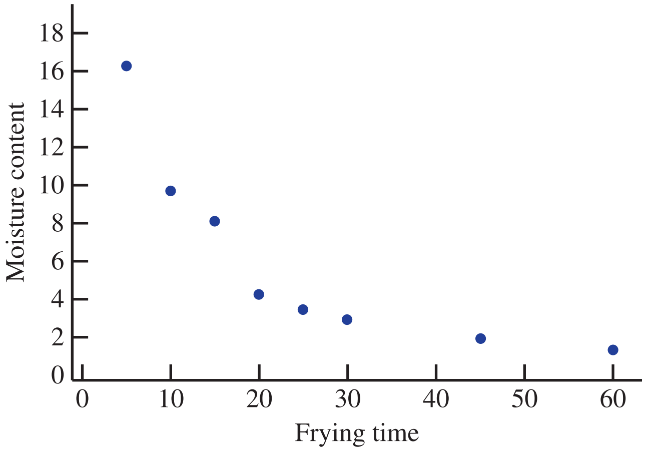


Yes, the pattern in the scatterplot is more nearly straight than the one in Part (a).

d. Taking the cube roots of both variables might be successful in straightening the plot. (In fact, this transformation does seem to result in a more nearly linear relationship, as shown in the scatterplot.)



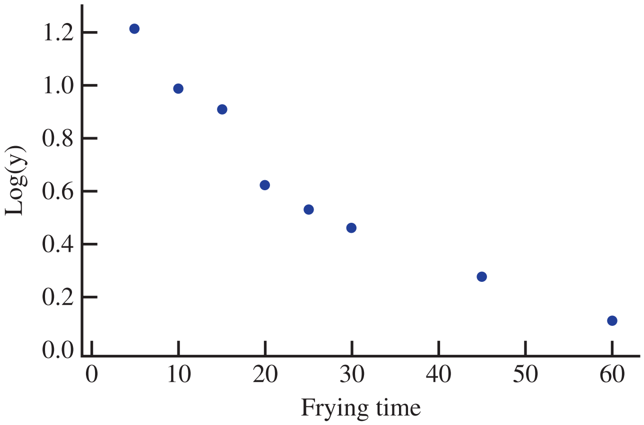
5.75 **a.**



No, the relationship does not seem to be linear.

Copyright 2025 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s). Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

b.



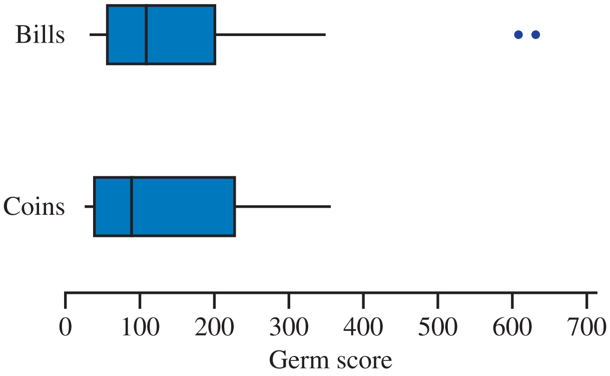
Yes, the transformed relationship is closer to being linear than the relationship shown in Part (a).
c. $y' = 1.143 - 0.0192x$, where x = frying time and y' is log(moisture content). **d.** 2.964%

5.77 **a.** The correlation coefficient of 0.12 indicates a weak relationship between price and weight of the wine bottles. **b.** $r^2 = 0.0144$. Approximately 1.44% of the variation in price of the wine can be explained by the linear relationship between price and weight of the wine bottles.

5.79 **a.** $r = 0$ **b.** For example, adding the point (6, 1) gives $r = 0.510$. (Any y-coordinate greater than 0.973 will work.) **c.** For example, adding the point (6, −1) gives $r = -0.510$. (Any y-coordinate less than −0.973 will work.)

Cumulative Review Exercises 5

CR5.3 **a.**



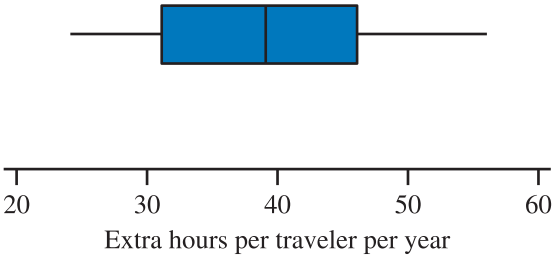
b. The distributions of germ score for both bills and coins are positively skewed. The median germ score for bills is greater than that for coins. The spread of the middle 50% of the data for coins is greater than that for bills. There are two outliers in germ scores for bills.

CR5.5 **a.** $r = 0.394$ **b.** $r = -0.664$ **c.** There is a stronger relationship between Happiness Index and response to Statement 2. **d.** There is a weak, positive association between the Happiness Index and the response to Statement 1. In contrast, there is a moderate, negative association between the Happiness Index and the response to Statement 2.

CR5.7 $r = 0.975$; this value is consistent with the previous answer, because the correlation coefficient is large and positive (close to 1), which indicates a strong positive association between the amount spent on science and the amount spent on pets.

CR5.9 Lower quartile = 7th value = 10478 mg/kg.
Upper quartile = 20th value = 11778 mg/kg.
Interquartile range = 11778 − 10478 = 1300 mg/kg.

CR5.11 **a.** The mean Extra Hours per Traveler per Year is 38.60, and the median is 39. Given that the mean and median values are nearly equal, we can conclude that the distribution is approximately symmetric.
b.



The distribution of extra hours per traveler per year is approximately symmetric, with a median value of 39. The IQR is approximately 15, and the values vary between a low of 24 to a high of 56. There are no outliers in the distribution.

CR5.13 **a.** The y intercept is −147. **b.** The slope is 6.175. For every 1-cm increase in the snout-vent length, the predicted clutch size increases by 6.175. **c.** It would be inadvisable to use the least-squares line to predict the clutch size when the snout-vent length is 22, since 22 is well outside the range of the x values in the original data set, and we have no reason to think that the observed linear relationship applies outside this range.

Chapter 6

6.1 A chance experiment is any activity or situation in which there is uncertainty about which of two or more possible outcomes will result.

6.3 **a.** {AA, AM, MA, MM}.

6.5 **a.** A = {Head oversize, Prince oversize, Slazenger oversize, Wimbledon oversize, Wilson oversize}
b. B = {Wimbledon midsize, Wilson midsize, Wimbledon oversize, Wilson oversize} **c.** not B = {Head midsize, Prince midsize, Slazenger midsize, Head oversize, Prince oversize, Slazenger oversize}

6.7 **b. i.** $A^C = \{(15, 8), (15, 16), (15, 32)\}$

ii. $A \cup B = \begin{Bmatrix} (10, 8), (10, 16), (10, 32), (12, 8) \\ (12, 16), (12, 32), (15, 16), (15, 32) \end{Bmatrix}$

iii. $A \cap B = \{(10, 16), (10, 32), (12, 16), (12, 32)\}$

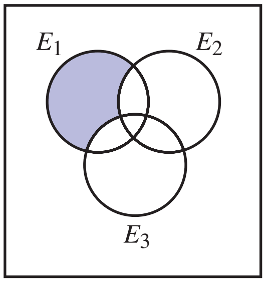
c. A and C are not disjoint. B and C are not disjoint.

6.9 **a.** $A = \{(\text{D1}), (\text{D2}), (\text{D3})\}$ **b.** $C = \{(\text{M1, M2, D1}), (\text{M1, M2, D2}), (\text{M1, M2, D3}), (\text{M2, M1, D1}), (\text{M2, M1, D2}), (\text{M2, M1, D3})\}$

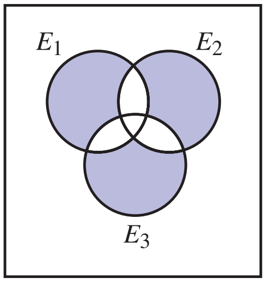
6.11 **a.** $A = \{\text{NN, DNN, NDN}\}$ **b.** $B = \{\text{DDNN, DNDN, NDDN}\}$ **c.** The number of outcomes is infinite.

6.13 **a.** $B^C = \{(1,1,1), (1,1,2), (1,1,3), (1,2,1), (1,2,2), (1,3,1), (1,3,3), (2,1,1), (2,1,2), (2,2,1), (2,2,2), (2,2,3), (2,3,2), (2,3,3), (3,1,1), (3,1,3), (3,2,2), (3,2,3), (3,3,1), (3,3,2), (3,3,3)\}$
b. $C^C = \{(1,1,2), (1,2,1), (1,2,2), (1,2,3), (1,3,2), (2,1,1), (2,1,2), (2,1,3), (2,2,1), (2,2,2), (2,2,3), (2,3,1), (2,3,2), (2,3,3), (3,1,2), (3,2,1), (3,2,2), (3,2,3), (3,3,2)\}$
c. $A \cup B = \{(1,1,1), (2,2,2), (3,3,3), (1,2,3), (1,3,2), (2,1,3), (2,3,1), (3,1,2), (3,2,1)\}$ **d.** $A \cap B = \varnothing$
e. $A \cap C = \{(1,1,1), (3,3,3)\}$

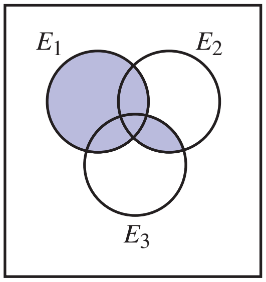
6.15 **a.**



b.



c.



6.17 {expedited overnight delivery, expedited second-business-day delivery, standard delivery, delivery to the nearest store for customer pick-up}

6.19 **a.** In the long run, about 0.1% of eggs are double yolk. **b.** We would expect to find about 5 double-yolk eggs because 0.1% of 5000 is 5.

6.21 **a.** {fiction hardcover, fiction paperback, fiction digital, fiction audio, nonfiction hardcover, nonfiction paperback, nonfiction digital, nonfiction audio} **b.** No

6.23 **a. i.** $A = \{(\text{C, N}), (\text{N, C}), (\text{N, N})\}$.
ii. $P(A) = 0.09 + 0.09 + 0.01 = \mathbf{0.19}$.
b. i. $B = \{(\text{C, C}), (\text{N, N})\}$. **ii.** $P(B) = 0.81 + 0.01 = \mathbf{0.82}$.

6.25 **a.** 0.225 **b.** 0.775 **c.** 0.06 **d.** 0.715 **e.** 0.234

6.27 **a.** 0.3 **b.** No. The problems to be graded will be selected at random, so the probability of being able to turn in both of the selected problems is the same whatever your choice of three problems to complete. **c.** 0.6

6.29 If the events F and I were mutually exclusive, we could add the probabilities to obtain $P(F \cup I)$. Since $P(F) + P(I) = 0.71 + 0.52 = 1.23$ is greater than 1 (and probabilities cannot be greater than 1), we know that F and I cannot be mutually exclusive.

6.31 **a.** 0.276 **b.** 0.49 **c.** 0.173

6.33 0.11

6.35 **a.** BC, BM, BP, BS, CM, CP, CS, MP, MS, PS
b. 0.1 **c.** 0.4 **d.** 0.3

6.37 **a.** $P(O_1) = P(O_3) = P(O_5) = 1/9$ and $P(O_2) = P(O_4) = P(O_6) = 2/9$ **b.** 1/3, 4/9 **c.** 9/21, 2/7

6.39 **a.** $P(\text{opioid} \mid \text{non-Hispanic Black}) = 0.08/0.23 = 0.35$ **b.** They are not equal to each other. The probability calculated in Part (a) is less than the unconditional probability that a patient visiting the emergency department for severe abdominal pain was given a prescription for an opioid pain medication. **c.** Given that the conditional probabilities are smaller than the unconditional probabilities, one can conclude that non-Hispanic Black or Hispanic patients visiting the emergency department are less likely to be given a prescription for an opioid pain medication.

6.41 **a. i.** 1005 respondents total, 302 have at least one tattoo: $P(T) = 302/1005 = 0.300$ **ii.** 237 respondents in West, 71 have at least one tattoo: $P(T \mid W) = 71/237 = 0.300$ **iii.** 179 respondents in Northeast, 54 have at least one tattoo: $P(T \mid N) = 54/179 = 0.302$
b. i. The probability that the selected person has at least one tattoo is 0.300. **ii.** The probability that the selected person has at least one tattoo given that the selected person lived in the West is 0.300. **iii.** The probability that the selected person has at least one tattoo given that the selected person lived in the Northeast is 0.302.
c. The computed probabilities are all very close to each other. As such, it could be argued that there are no significant regional differences in the likelihood that individuals have at least one tattoo.

6.43 **a.** 0.440 **b.** 0.560 **c.** 0.570 **d.** 0.260

6.45 **a.** 0.1945 **b.** 0.0749 **c.** 0.0616 **d.** 0.4777

6.47 **a.** 0.769 **b.** 0.890 **c.** Since the conditional probabilities in (a) and (b) are not equal, a prediction that a baby is male and a prediction that a baby is female are not equally reliable.

6.49 **a.** $P(\text{wear seat belt} \mid \text{urban area}) = 1799/(1799 + 201) = 1799/2000 = 0.8995$ **b.** $P(\text{wear seat belt} \mid$

Copyright 2025 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s). Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

rural area) = 1607/(1607 + 393) = 1607/2000 = 0.8035 **c.** The probability is higher for those in urban areas, so those are the drivers that are more likely to wear seat belts.

6.51 **a.** 0.31 **b.** 0.43 **c.** 0.23

6.53 **a.** Answers will vary. One example is to let event *A* be selecting a sophomore, and event *B* be selecting a senior. **b.** Answers will vary. One example is to let event *C* be selecting a member of the soccer team, and event *D* be selecting a junior.

6.55 **a.** 85% **b.** 0.15

6.57 **a.** Given that the selected graduate finished college with no student debt, the probability that the selected graduate strongly agrees that education was worth the cost is 0.49. **b.** Given that the selected graduate finished college with high student debt, the probability that the selected graduate strongly agrees that education was worth the cost is 0.18. **c.** We are given that $P(A) = 0.38$ and $P(A|H) = 0.18$. The events *A* and *H* are not independent because $P(A|H) \neq P(A)$.

6.59 **a.** 0.001. We have to assume that they deal with the three errands independently. **b.** 0.999 **c.** 0.009

6.61 **a.** 0.9931, increases **b.** 0.9266, decreases

6.63 The reason that $P(T)$ is not the average of the two given conditional probabilities is because there are different numbers (or different proportions) of people in the two given age groups (19 to 36 and 37 or older).

6.65 **a.** 0.336 **b.** 0.56 **c.** 0.06 **d.** 0.18

6.67 **a.** 0.00391 **b.** 0.00383

6.69 **a.** 0.5625 **b.** 0.34 **c.** 0.105

6.71 **a.** 0.55 **b.** 0.45 **c.** 0.4 **d.** 0.25

6.73 **a.** 0.032. **b.** 0.032

6.75 **a.** 0.49 **b.** 0.3125 **c.** 0.24 **d.** 0.1

6.77 **a.** $P(\textit{not } C) = 1 - P(C) = 1 - 0.04 = 0.96$ **b.** $P(+ \mid \textit{not } C) = 0.13$ (false positive) **c.** $P(- \mid \textit{not } C) = 1 - 0.13 = 0.87$ **d.** $P(- \mid C) = 0.08$ (false negative) **e.** $P(+ \mid C) = 1 - 0.08 = 0.92$

f. $$P(C \mid +) = \frac{P(+\mid C)P(C)}{P(+\mid C)P(C) + P(+\mid \textit{not } C)P(\textit{not } C)}$$

$$= \frac{(0.92)(0.04)}{(0.92)(0.04) + (0.13)(0.96)} = 0.228$$

g. $P(+ \mid C)$ is the probability that the randomly selected person tests positive given that they have colon cancer. $P(C \mid +)$ is the probability that the randomly selected person has colon cancer given that they test positive.

6.79 **a.** *P*(long COVID | did not complete high school) = 0.409. **b.** The given conditional probabilities indicate that higher levels of educational attainment are associated with lower proportions of long COVID. Those without a high school degree have the highest probability of experiencing long COVID, and those with a college degree have the lowest probability of experiencing long COVID.

6.81 **a.** 0.4636 **b.** 0.1598 **c.** The probability in Part (a) is the probability of the selected student having no student loan debt given that the selected student strongly agrees that education was worth the cost, which is 0.4636. The probability in Part (b) is the probability of the selected student having student loan debt of more than $50,000 given that the selected student strongly agrees that education was worth the cost, which is 0.0598.

6.83 The reason that $P(C)$ is not the average of the three conditional probabilities is that there are different numbers of people driving the three different types of vehicles (and that there are some drivers who are driving vehicles not included in those three types).

6.85

	Radiologist 1		
	Predicted Male	**Predicted Female**	**Total**
Baby Is Male	74	12	**86**
Baby Is Female	14	59	**73**
Total	**88**	**71**	**159**

a. *P*(prediction is male | baby is male) = 74/86 = 0.860. **b.** *P*(prediction is female | baby is female) = 59/73 = 0.808. **c.** Yes. Since the answer to (a) is greater than the answer to (b), the prediction is more likely to be correct when the baby is male. **d.** The radiologist was correct for 74 + 59 = 133 of the 159 babies in the study. So *P*(correct) = 133/159 = 0.836.

6.87 **a.** **i.** 0.99 **ii.** 0.01 **iii.** 0.99 **iv.** 0.01 **b.** 0.02 **c.** 0.5, it is consistent with the quote.

6.89 **b.** 0.27 **c.** 0.43 **d.** Not independent

6.91 **a.** 0.3700 **b.** 0.1674 **c.** 0.2923 **d.** 0.2900

6.93 **a.** 0.7493 **b.** 0.3455 **c.** 0.4284 **d.** 0.1193

6.101 **b.** 0.025 **c.** 0.069

6.103 **a.** 0.49 **b.** 0.88 **c.** 0.7

6.105 The events *selected adult exercises at least 30 minutes 3 times per week* and *is a millennial* are dependent events. A millennial is more likely to exercise at least 30 minutes 3 times per week (57.1%) when compared with adults over age 35 (51.1%). The age of the adult (millennial or over age 35) has an impact on the length of time the selected adult exercises.

6.107 **a.** $P(D) = 0.148$, $P(T) = 0.374$, and $P(D \cap T) = 0.106$ **b.** $P(D \cup T) = 0.416$. *P*(neither a change in diagnosis nor change in treatment) = $1 - P(D \cup T) = 0.584$ **c.** 0.416

6.109 **a.** 0.018 **b.** $P(A_1|L) = 0.444$ **c.** $P(A_2|L) = 0.278$ **d.** $P(A_3|L) = 0.278$

6.111 **a.** For example, use a single-digit random number to represent the outcome of the game. The digits 0–7 will represent a win for seed 1, and digits 8–9 will represent a win for seed 4. **b.** For example, use a single-digit random number to represent the outcome of the game. The digits 0–5 will represent a win for seed 2, and digits 6–9 will represent a win for seed 3. **c.** For example, use a single-digit random number to represent the outcome of the game. If seed 1 won game 1 and seed 2 won game 2, the digits 0–5 will represent a win for seed 1, and digits 6–9 will represent a win for seed 2. If seed 1 won game 1 and seed 3 won game 2, the digits 0–6 will represent a win for seed 1, and digits 7–9 will represent a win for seed 3. If seed 4 won game 1 and seed 2 won game 2, the digits 0–6 will represent a win for seed 2, and digits 7–9 will represent a win for seed 4. If seed 4 won game 1 and seed 3 won game 2, the digits 0–5 will represent a win for seed 3, and digits 6–9 will represent a win for seed 4. **d.** Answers will vary.

6.113 **a.** 0.4 **b.** 0.75 **c.** 0.25

6.115 0.49

6.117 Since the total number of viewers was 2517 and the number of viewers of R-rated movies was 1140, $P(R_2|R_1) = 1139/2516$ and $P(R_2|R_1^C) = 1140/2516$. These probabilities are not equal, so the events R_1 and R_2 are not independent. However, as a result of the fact that the total number of viewers is large, the two probabilities are very close, and therefore from a practical point of view the events may be regarded as independent.

6.119 **a.** 0.383 **b.** 0.617

Chapter 7

7.1 **a.** Discrete **b.** Continuous **c.** Discrete **d.** Discrete **e.** Continuous

7.3 **a.** Positive integers **b.** For example, S with $y = 1$, RS, LS with $y = 2$, RLS, RRS, LES, LLS with $y = 2$, RLS, RRS, LES, LLS with $y = 3$.

7.5 **a.** The possible values are the real numbers between 0 and 100, inclusive; **b.** y is continuous.

7.7 **a.** 3, 4, 5, 6, 7 **b.** −3, −2, −1, 0, 1, 2, 3 **c.** 0, 1, 2 **d.** 0, 1

7.9 $P(3 \leq x \leq 6) = 0.09 + 0.25 + 0.4 + 0.16 = 0.9$. $P(3 < x < 6) = 0.25 + 0.4 = 0.65$. The second probability is smaller (and therefore the two probabilities are not equal) because the second probability does not include the possibility of a student taking three or six courses, while the first does include this possibility.

7.11 **a.** 0.1 **b.** 0.95

7.13 **a.** 0.82 **b.** 0.18 **c.** 0.65, 0.27

7.15 Results of the simulation will vary.

7.17 **a.**

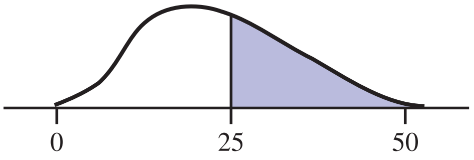
x	0	1	2	3	4
$p(x)$	0.4096	0.4096	0.1536	0.0256	0.0016

b. 0 and 1 **c.** 0.1808

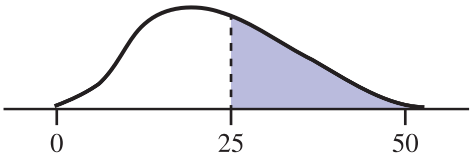
7.19 **a.** The smallest possible y value is 1, and the corresponding outcome is S. The second smallest y value is 2, and the corresponding outcome is FS. **b.** The set of positive integers.
c. $p(y) = (0.3)^{y-1}(0.7)$, for y = 1, 2, 3, . . .

7.21 **a.** 2, 3, 4, … **b.** 0.64 **c.** 0.256

7.23 **a.**



b.



7.25
$P(2 < x < 3) = P(2 \leq x \leq 3) = (1/10)(3 - 2) = 0.1$
$P(x < 2) = (1/10)(2 - 0) = 0.2$
$P(x > 7) = (1/10)(10 - 7) = 0.3$
Thus,
$P(2 < x < 3) = P(2 \leq x \leq 3) < P(x < 2) < P(x > 7)$

7.27 **a.** 0.25 **b.** 0.25 **c.** 0.25 **d.** These probabilities are the same because the density curve is uniform (always the same height) and the x-values of interest span the same range (the change in x is 100 in both cases). The rectangular regions under the density curve are the same size.

7.29 **a.** $P\left(x > \frac{1}{2}\right) = \frac{1}{4}$

b. $P\left(x \geq \frac{1}{4}\right) = \frac{9}{16}$

7.31 **a.** Area $= \frac{1}{2}(40)(0.05) = 1$, as required.

b. $P(w < 20) = 0.5$

c. $P(w < 10) = 0.125$

d. $P(w > 30) = 0.125$

7.33 **a.** 2.4317, 1.3797 **b.** The mean calculated in Part (a) is likely less than the actual mean household size in 2021 because there are some households with more than 7 individuals, which will pull the mean up. **c.** The reported mean household of size 2.54 is greater than that calculated in Part (a). This supports the answer in Part (b).

7.35 **a.** 0.56 **b.** 65% **c.** The mean is not $(0 + 1 + 2 + 3 + 4)/5$ since, for example, far more cartons have 0 broken eggs than 4 broken eggs, and so 0 needs a much greater weighting than 4 in the calculation of the mean.

7.37 **a.** 0.14 **b.** $\mu - 2\sigma = 4.66 - 2(1.2) = 2.26$. $\mu + 2\sigma = 4.66 + 2(1.2) = 7.06$. The values of x more than 2 standard deviations away from the mean are 1 and 2. **c.** The probability that x is more than 2 standard deviations away from its mean is 0.05.

7.39 **a.** 2.8, 1.288 **b.** 0.7, 0.781

7.41 **a.** 3.5, 2.27 **b.** 15.4, 75.94

7.43 **a.** 46.5 **b.** $890

7.47 **a.** Whether y is positive or negative tells us whether or not the peg will fit into the hole. **b.** 0.003 **c.** 0.00632 **d.** Yes. Since there is no reason to believe that the pegs are being specially selected to match the holes (or vice versa), it is reasonable to think that x_1 and x_2 are independent. **e.** Since 0 is less than half a standard deviation from the mean in the distribution of y, it is relatively likely that a value of y will be negative, and, therefore, that the peg will be too big to fit the hole.

7.49 **a.** 3.5, 2.9167, 1.7078 **b.** 3.5, 2.9167, 1.7078

7.51 **a.** 0.9938 **b.** 0.9996 **c.** 0.0004 **d.** 0.3347

7.53 **a.** 0.001 **b.** 0.001 **c.** 0.147

7.55 **a.** 0.2335 **b.** 0.6172 **c.** 0.8497

7.57 **a.** 0.2573 **b.** The computed probabilities are shown below. It is somewhat more likely that 2 or fewer use the same password for all online accounts when compared with more than 2 using the same password.

$$P(x \le 2) = p(0) + p(1) + p(2) = 0.5080$$
$$P(x > 2) = 1 - P(x \le 2) = 1 - 0.5080 = 0.4920$$

7.59 **a.** The probability distribution of x is geometric, with success probability $p = 0.49$. The distribution is geometric because we are waiting until we find someone that washes sheets at least once a week. **b.** 0.1274 **c.** 0.8673 **d.** 0.1327

7.61 **a.** 200 **b.** 13.416

7.63 **a.** Binomial distribution with $n = 100$ and $p = 0.2$ **b.** 20 **c.** 16, 4 **d.** A score of 50 is $(50 - 20)/4 = 7.5$ standard deviations from the mean in the distribution of x. So, a score of over 50 is very unlikely.

7.65 **a.** 0.017 **b.** 0.811, **c.** 0.425 **d.** The error probability when $p = 0.7$ is now 0.902. The error probability when $p = 0.6$ is now 0.586.

7.67 **a.** There is not a fixed number of trials. **b. i.** 0.062 **ii.** 0.284 **iii.** 0.716 **iv.** 0.779

7.69 **a.** 0.0975 **b.** 0.043 **c.** 0.815

7.71 **a.** 0.1003 **b.** 0.1003 **c.** 0.8185 **d.** 0.5 **e.** 1.0000 **f.** 0.939 **g.** 0.5910

7.73 **a.** 0.0706 **b.** 0.0228 **c.** 0.9996 **d.** approximately 1

7.75 **a.** $z^* = 1.88$ **b.** $z^* = 2.33$ **c.** $z^* = -1.75$ **d.** $z^* = -1.28$

7.77 **a.** $z^* = 1.34$ **b.** $z^* = 0.74$ **c.** $z^* = 0$ **d.** $z^* = -1.34$ **e.** The value of the 30th percentile is the negative of the value of the 70th percentile.

7.79 **a.** 0.02275 **b.** 0.656 **c.** 0.656

7.81 7.3%

7.83 **a.** 0.2033 **b.** 0.5934 **c.** 0.0124

7.85 Those with emissions > 2.113 parts per billion

7.87 **a.** 0.2843 **b.** 0.0918 **c.** 0.4344 **d.** 29.928

7.89 **a.** 0.0035 **b.** 0.382

7.91 The second machine

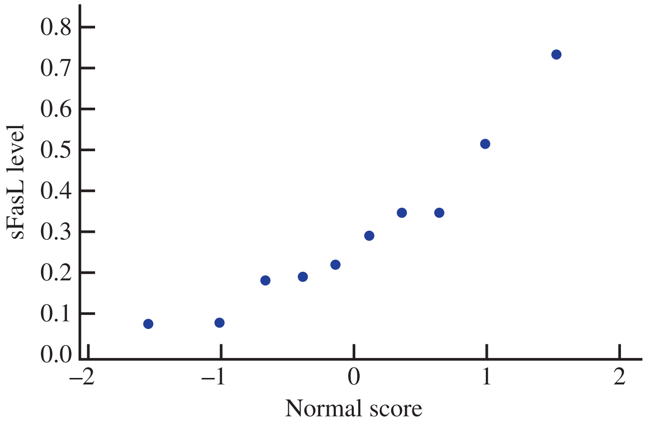
7.93 **a.** 0.5953 **b.** Yes, $P(x < 25) = 0.000015$ **c.** 0.00011 **d.** 28.026

7.95 **a.** 0.9876 **b.** 0.9888 **c.** 6.608

7.97 Of the three plots given, the second one is the only one that is reasonable to think that the population distribution is approximately normal because it is the only plot with an approximately linear pattern of points. The two other plots show clear curvature.

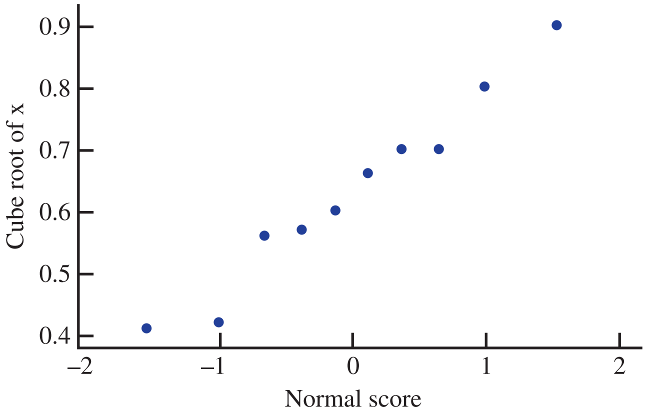
7.99 **b.** No. The distribution of x is positively skewed. **d.** Yes. The histogram shows a distribution that is slightly closer to being symmetric than the distribution of the untransformed data. **f.** Both transformations produce histograms that are closer to being symmetric than the histogram of the untransformed data, but neither transformation produces a distribution that is truly close to being normal.

7.101 **a.**



b. The normal probability plot appears curved.

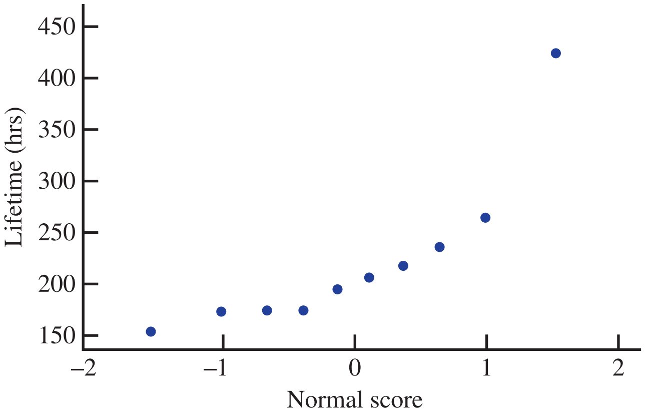
c.



Yes. This normal probability plot appears more linear than the plot for the untransformed data.

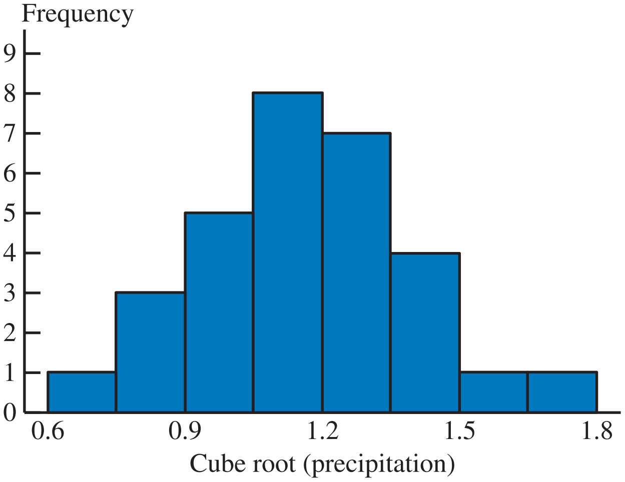
Copyright 2025 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s). Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

7.103



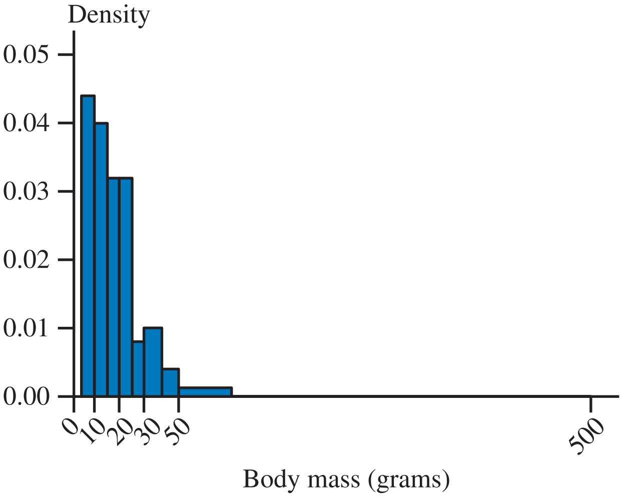
Since the normal probability plot shows a clear curve, the normal distribution is not a plausible model for power supply lifetime. However, it is worth noting that most of the apparent curved pattern is brought about by the single point with coordinates (1.539, 422.6).

7.105



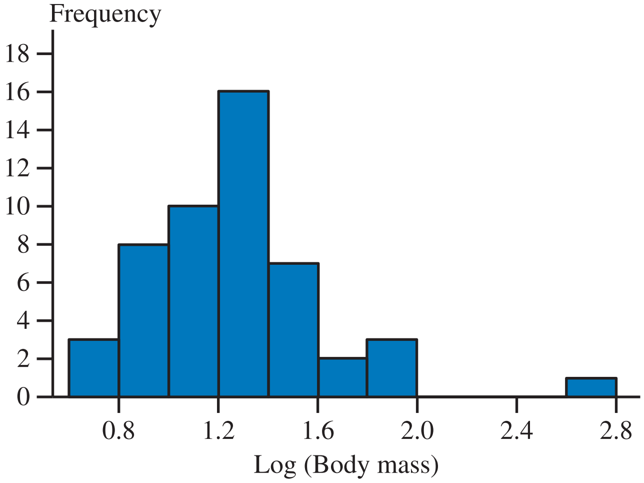
The cube-root transformation appears to result in the more symmetrical histogram.

7.107 **a.**



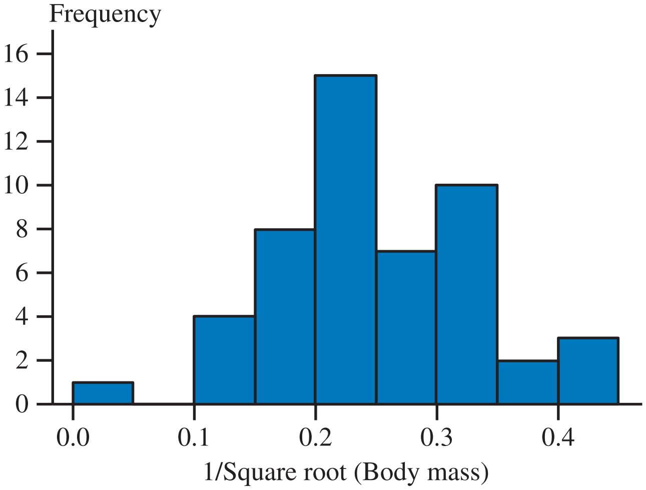
Yes. If a symmetrical distribution is required, a transformation is desirable.

b.



The distribution is closer to being symmetrical than the original distribution but is still positively skewed.

c.



The shape in the histogram roughly resembles that of a normal curve. Certainly, this transformation was more successful than the one in Part (b) in producing an approximately normal distribution.

7.109 **a.** 0.026 **b.** 0.7580 **c.** 0.7366 **d.** 0.9109

7.111 0.3084

7.113 **a.** 0.1114 **b.** 0.0409 **c.** 0.0968 **d.** 0.9429 **e.** 0.9001

7.115 **a.** When $p = 0.96$, $n(1 - p) = 60(0.04) = 2.4$, which is less than 10, and so the normal approximation to the binomial distribution should not be used. The binomial formula should be used. **b.** For a person who is not faking the test, a score of 42 or less is extremely unlikely (a probability of 0.000000000013). Therefore, if someone does get a score of 42 or less, we would be convinced that the person is faking the test.

7.117 **a.** 0.7970 **b.** 0.7016 **c.** 0.0143 **d.** 0.05

Copyright 2025 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s). Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

7.119 **a.** No, since $np = 50(0.05) = 2.5 < 10$. **b.** Now $n = 500$ and $p = 0.05$, so $np = 500(0.05) = 25 \geq 10$. So the techniques of this section can be used $P(\text{at least 20 are defective}) \approx 0.8708$.

7.121 **a.** 0.8245 **b.** 0.0521 **c.** 0.6826 **d.** 0.3783

7.123 **a.** 0.7 **b.** 0.45 **c.** 0.55

7.125 **a.** $\mu_x = 2.64$ $\sigma_x^2 = 2.3704$. $\sigma_x = 1.540$
b. $\mu_x - 3\sigma_x = 2.64 - 3(1.540) = -1.98$
$\mu_x + 3\sigma_x = 2.64 + 3(1.540) = 7.26$
Since all the possible values of x lie between these two values, the required probability is 0.

7.127 **a.** 0.25 **b.** 0.25 **c.** 18 minutes

7.129 **a.** If the coin is fair, then the distribution of x is binomial with $n = 25$ and $p = 0.5$. The probability that it is judged to be biased is $P(x \leq 7) + P(x \geq 18)$. $P(x \leq 7) = 0.014 + 0.005 + 0.002 = 0.021$. $P(x \geq 18) = 0.014 + 0.005 + 0.002 = 0.021$. So the probability that the coin is judged to be biased is $0.021 + 0.021 = 0.042$. **b.** Now the distribution of x is binomial with $n = 25$ and $p = 0.9$. The probability that the coin is judged to be biased is $P(x \leq 7) + P(x \geq 18)$. $P(x \leq 7) = 0.000$. $P(x \geq 18) = 0.997$. So the probability that the coin is judged to be fair is $1 - 0.997 = 0.003$. When $P(H) = 0.1$, the distribution of x is binomial with $n = 25$ and $p = 0.1$. The probability that the coin is judged to be biased is $P(x \leq 7) + P(x \geq 18)$. $P(x \leq 7) = 0.998$. $P(x \geq 18) = 0.000$. So the probability that the coin is judged to be fair is $1 - 0.998 = 0.002$. **c.** When $P(H) = 0.6$, the distribution of x is binomial with $n = 25$ and $p = 0.6$. The probability that the coin is judged to be biased is $P(x \leq 7) + P(x \geq 18)$. $P(x \leq 7) = 0.001$. $P(x \geq 18) = 0.153$. So the probability that the coin is judged to be fair is $1 - 0.153 = 0.847$. By symmetry, when $P(H) = 0.4$, the probability that the coin is judged to be fair is 0.847. These probabilities are larger than those in Part (b), since, when $P(H) = 0.6$ or $P(H) = 0.4$, the probability of getting a head is closer to 0.5 than when $P(H) = 0.9$ or $P(H) = 0.1$. Thus it is more likely that the coin will be judged to be fair when $P(H) = 0.6$ or $P(H) = 0.4$. **d.** It is now more likely that the coin will be judged to be fair, and so the error probabilities are increased. This would seem to make the rule less good; however, this rule makes it more likely that the coin will be judged to be fair when in fact it *is* fair.

7.131 **a.** 0.9332 **b.** 72.8 minutes **c.** \$60

7.133 **a.**

y	0	1	2	3	4
$p(y)$	0.0625	0.4375	0.3125	0.1250	0.0625

$\mu_y = 1.6875$

b.

y	0	1	2	3	4
$p(y)$	0.0256	0.3264	0.3456	0.1728	0.1296

$\mu_y = 2.0544$

c.

z	1	2	3	4
$p(z)$	0.1250	0.5000	0.2500	0.1250

$\mu_z = 2.375$

7.135 **a.** 0.1552 **b.** 0.2688
c.

x	4	5	6	7
$p(x)$	0.1552	0.2688	0.29952	0.27648

d. 5.697

7.137 **a.** 0.7745 **b.** 0.1587 **c.** 0.3085 **d.** 6.1645

7.139 **a.** No. The proportion of the population that is being sampled is $5000/40000 = 0.125$, which is more than 5%. **b.** $\mu = np = 100(0.275) = \mathbf{27.5}$ and $\sigma = \sqrt{np(1-p)} = 4.465$. **c.** No. Since n is being doubled, the standard deviation, $\sqrt{np(1-p)}$, is multiplied by $\sqrt{2}$.

Cumulative Review Exercises 7

CR7.3 No. The percentages given in the graph are said to be, for each year, the "percent increase in the number of communities installing" red-light cameras. This presumably means the percentage increase in the number of communities with red-light cameras *installed*, in which case the positive results for all of the years 2003 to 2009 show that a great many more communities had red-light cameras installed in 2009 than in 2002.

CR7.5 0.1243

CR7.7 $P(\text{Service 1}|\text{Late}) = 0.349$ and $P(\text{Service 2}|\text{Late}) = 0.651$. Service 2 is more likely to have been used.

CR7.9 **a.** Supplier 1 **b.** Supplier 2 **c.** I would recommend supplier 1 because the bulbs will last, on average, longer than those from supplier 2. Additionally, the bulbs from supplier 1 have less variability in their lifetimes, so there is more consistency in the bulb lifetimes when compared with supplier 2. **d.** Approximately 1000 hours. **e.** Approximately 100 hours.

CR7.11 **a.** 0.09 **b.** 0.045 **c.** 0.135 **d.** 0.667

CR7.13 **a.** $\mu_x = 2.3$. **b.** $\sigma_x^2 = 0.81$, $\sigma_x = 0.9$

CR7.15 **a.** 0.1587 **b.** 51.4 minutes **c.** 41.65 minutes

CR7.17 **b.** Positive **c.** Yes. If a symmetrical distribution were required, then a transformation would be advisable.

Chapter 8

8.1 A population characteristic is a quantity that summarizes the whole population. A statistic is a quantity calculated from the values in a sample.

8.3 **a.** Population characteristic **b.** Statistic **c.** Population characteristic **d.** Statistic **e.** Statistic

8.5 **a.** The histogram on the left indicates that the value of $\hat{p}$ has more sample-to-sample variability than the histogram on the right. Although the two histograms are similar, the histogram on the left has values that tend to deviate more from the center than the histogram on the right, indicating a larger standard deviation of the values of $\hat{p}$. **b.** Samples of size $n = 75$ would yield a value of $\hat{p}$ that is more likely to be close to 0.55. The fact that the histogram on the right seems to have less sample-to-sample variability, is centered at 0.55, and has tall bars close to 0.55 supports this choice.

8.7 **a.** and **b.**

$\bar{x}$	1.5	2	2.5	3	3.5
$p(\bar{x})$	1/6	1/6	1/3	1/6	1/6

c.

$\bar{x}$	1	1.5	2	2.5	3	3.5	4
$p(\bar{x})$	1/16	1/8	3/16	1/4	3/16	1/8	1/16

d. Both distributions are symmetrical, and their means are equal (2.5). However, the "with replacement" version has a greater spread than the first distribution, with values ranging from 1 to 4 in the "with replacement" distribution and from 1.5 to 3.5 in the "without replacement" distribution. The stepped pattern of the "with replacement" distribution more closely resembles a normal distribution than does the shape of the "without replacement" distribution.

8.9 **a.** and **b.**

$\bar{x}$	$2^2/_3$	3	$3^1/_3$	$3^2/_3$
$p(\bar{x})$	0.1	0.4	0.3	0.2

Sample Median	3	4
p(Sample Median)	0.7	0.3

(Max + Min)/2	2.5	3	3.5
p((Max + Min)/2)	0.1	0.5	0.4

c. Since $\mu = 3.2$ and $\mu_{\bar{x}} = 3.2$, $\bar{x}$ is an unbiased estimator of μ, which is not the case for either of the two other statistics. Since the distribution of the sample mean has less variability than either of the other two sampling distributions, the sample mean will tend to produce values that are *closer* to μ than the values produced by either of the other statistics.

8.11 The sampling distribution of $\bar{x}$ will be approximately normal for the sample sizes in Parts (c)–(f), since those sample sizes are all greater than or equal to 30.

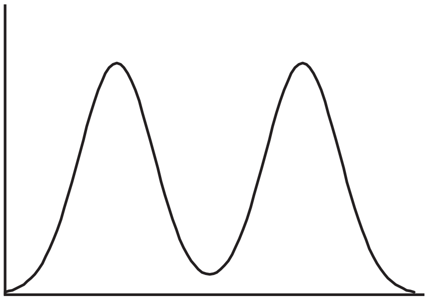
8.13 **a.** 0.237 **b.** Yes, because $n = 30$.

8.15 **a.** $\mu_{\bar{x}} = 40$, $\sigma_{\bar{x}} = 0.625$ approximately normal **b.** 0.5762 **c.** 0.2628

8.17 **a.** $\mu_{\bar{x}} = 2$, $\sigma_{\bar{x}} = 0.267$ **b.** In each case $\mu_{\bar{x}} = 2$. When $n = 20$, $\sigma_{\bar{x}} = 0.179$, and when $n = 100$, $\sigma_{\bar{x}} = 0.008$. **c.** All three centers are the same, and the larger the sample size, the smaller the standard deviation of $\bar{x}$. Since the distribution of $\bar{x}$ when $n = 100$ is the one with the smallest standard deviation of the three, this sample size is most likely to result in a value of $\bar{x}$ close to μ.

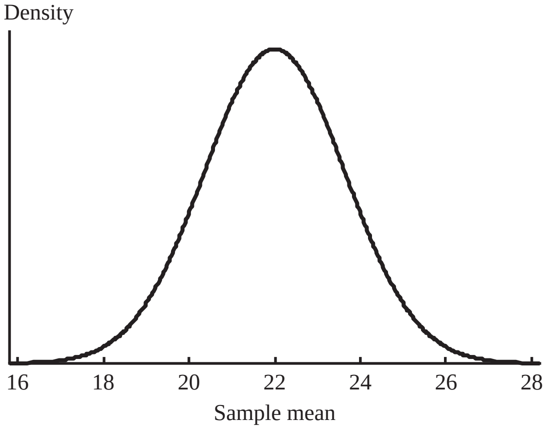
8.19 **a.** 0.6826 **b. i.** 2 **ii.** 3

8.21 **a.** In the sketch given below, the horizontal axis represents x (the number of checks written) and the vertical axis represents density.



b. $\mu_{\bar{x}} = \mu = 22$, and $\sigma_{\bar{x}} = \dfrac{\sigma}{\sqrt{n}} = \dfrac{16.5}{\sqrt{100}} = 1.65$

Since $n = 100 \geq 30$, the distribution of $\bar{x}$ is approximately normal.



c. $P(\bar{x} \leq 20) \approx 0.1131$
$P(x \geq 25) \approx 0.0344$

8.23 **a.** $\mu_{\hat{p}} = 0.65$, $\sigma_{\hat{p}} = 0.151$. **b.** $\mu_{\hat{p}} = 0.65$, $\sigma_{\hat{p}} = 0.107$. **c.** $\mu_{\hat{p}} = 0.65$, $\sigma_{\hat{p}} = 0.087$. **d.** $\mu_{\hat{p}} = 0.65$, $\sigma_{\hat{p}} = 0.067$. **e.** $\mu_{\hat{p}} = 0.65$, $\sigma_{\hat{p}} = 0.048$. **f.** $\mu_{\hat{p}} = 0.65$, $\sigma_{\hat{p}} = 0.034$.

8.25 **a.** The histogram on the left indicates that the value of $\hat{p}$ has more sample-to-sample variability because there is more spread in the values of $\hat{p}$ in the histogram. **b.** Samples of size $n = 110$ would have values of $\hat{p}$ that are more likely to be close to the actual population proportion of $p = 0.17$ because there is less variability in the histogram on the right.

8.27 The sampling distribution of the sample proportion is approximately normal with mean 0.09 and standard deviation 0.009316. Then $P(\hat{p} \geq 0.096) \approx 0.260$. It is not unlikely that a sample proportion as large as 0.096 would be observed if the true value was 0.09. Therefore, we do not have convincing evidence that the proportion of all recent college graduates who believe that Judith Sheindlin is a member of the U.S. Supreme Court is greater than 0.09.

Copyright 2025 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s). Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

8.29 **a.** If $p = 0.8$, then $\mu_{\hat{p}} = 0.8$ and

$$\sigma_{\hat{p}} = \sqrt{\frac{0.8(1 - 0.8)}{225}} = 0.027.$$

b. If $p = 0.7$, then $\mu_{\hat{p}} = 0.7$ and

$$\sigma_{\hat{p}} = \sqrt{\frac{0.7(1 - 0.7)}{225}} = 0.031.$$

c. When $p = 0.8$, $np = 225(0.8) = 180$ and $n(1 - p) = 225(1 - 0.8) = 45$, which are both greater than 10. When $p = 0.7$, $np = 225(0.7) = 157.5$ and $n(1 - p) = 225(1 - 0.7) = 67.5$, which are also both greater than 10. In each case, np and $n(1 - p)$ are both greater than 10, so the sampling distribution of $\hat{p}$ is approximately normal for $p = 0.8$ as well as $p = 0.7$.

8.31 The sampling distribution of the sample proportion is approximately normal with mean 0.5 and standard deviation 0.025. Then $P(\hat{p} \geq 0.51) \approx 0.345$. It is reasonably likely that a sample proportion as large as 0.51 would be observed if the true value were 0.5 or less. Therefore, it is not reasonable to conclude that the proportion of workers who say that seeing a coworker's dog in a Zoom meeting makes the meeting more enjoyable for them is greater than 0.5.

8.33 **a.** $\bar{x}$ is approximately normally distributed with mean 50 and standard deviation 0.1. **b.** 0.9876 **c.** 0.5

8.35 **a.** 0.294 **b.** Yes, it is reasonable to think that this estimate is within 0.05 of the actual value of the population proportion. The sampling distribution of the sample proportion is centered at p. The standard deviation of the sampling distribution of $\hat{p}$ is 0.0144, and the sampling distribution is approximately normal. Since the sampling distribution of $\hat{p}$ is approximately normal and is centered at the actual population proportion p, we now know that about 95% of all possible random samples of size $n =$ 1,005 will produce a sample proportion that is within $2(0.0144) = 0.0288$ of the actual value of the population proportion. This tells us that the sample estimate $\hat{p} = 0.294$ is likely to be within 0.0288 of the actual proportion of all adult Americans with at least one tattoo. This margin of error of 0.0288 is less than the value of 0.05, so it is reasonable to think that this estimate is within 0.05 of the actual value of the population proportion.

8.37 0.0793

Chapter 9

9.1 Statistics II and III are preferable to Statistic I since they are unbiased (their means are equal to the value of the population characteristic). However, Statistic II is preferable to Statistic III since its standard deviation is smaller. So Statistic II should be recommended.

9.3 $\hat{p} = \frac{584}{1{,}668} = 0.350$

9.5 $\hat{p} = 0.7$

9.7 **a.** $\bar{x} = 421.429$ **b.** $s^2 = 10414.286$ **c.** $s = 102.050$. No, s is not an unbiased statistic for estimating σ.

9.9 **a.** $\bar{x} = 120.6$ therms **b.** The value of τ is estimated to be $10000(120.6) = 1{,}206{,}000$ therms. **c.** $\hat{p} = 0.8$ **d.** sample median = 120 therms

9.11 The sample of size $n = 100$ is likely to result in a wider confidence interval. The formula for the confidence interval is $\hat{p} \pm 1.96\sqrt{\frac{\hat{p}(1 - \hat{p})}{n}}$, and so a smaller value of n is likely to result in a wider confidence interval since n is in the denominator.

9.13 **a.** Yes, since $np = 50(0.3) = 15 \geq 10$ and $n(1 - p) = 50(0.7) = 35 \geq 10$. **b.** No, since $np = 50(0.05) = 2.5$, which is not greater than or equal to 10. **c.** No, since $np = 15(0.45) = 6.75$, which is not greater than or equal to 10. **d.** No, since $np = 100(0.01) = 1$, which is not greater than or equal to 10.

9.15 **a.** The larger the confidence level, the wider the interval. **b.** The larger the sample size, the narrower the interval. **c.** Values of $\hat{p}$ farther from 0.5 give smaller values of $\hat{p}(1 - \hat{p})$. Therefore, the farther the value of $\hat{p}$ from 0.5, the narrower the interval.

9.17 If a large number of random samples of size 1000 were to be taken, about 90% of the resulting confidence intervals would contain the true proportion of all adult Americans who think that they could compete in the Olympics.

9.19 **a.** (0.548, 0.592). We are 95% confident that the actual proportion of all adult Americans who would give themselves a grade of A or B on their financial knowledge of personal finance is between 0.548 and 0.592. The method used to construct this interval estimate is successful in capturing the actual value of the proportion about 95% of the time. **b.** Because this confidence interval is entirely above 0.5, the interval is consistent with the statement that a majority of adult Americans would give themselves a grade of A or B.

9.21 **a.** (0.625, 0.675). We are 90% confident that the proportion of dog owners who take more pictures of their dog than of their significant others or friends is between 0.625 and 0.675. **b.** (0.429, 0.491). We are 95% confident that the actual proportion of dog owners who are more likely to complain to their dog than to a friend is between 0.429 and 0.491. **c.** First, the confidence interval in Part (b) has a higher confidence level, which uses a larger z critical value in the computation of the confidence interval. Second, the standard error in Part (b) is larger than that in Part (a). Both of these reasons contribute to the wider confidence interval in Part (b).

9.23 (0.193, 0.227). We are 95% confident that the proportion of working adults who are very concerned that their job will be automated, outsourced, or otherwise made obsolete in the next 5 years is between 0.193 and 0.227.

Copyright 2025 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s). Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

9.25 Intervals constructed when the sample proportions ($\hat{p}$) are closer to 0.5 are wider than those farther from 0.5 because the product $\hat{p}(1-\hat{p})$ is largest when $\hat{p} = 0.5$. In the previous exercise, $\hat{p} = 0.850$ and in this exercise, $\hat{p} = 0.810$. The sample proportion in this exercise is closer to 0.5 than that in the previous exercise. Therefore, the interval in this exercise will be wider than the interval in the previous exercise.

9.27 (0.362, 0.418). We are 90% confident that the proportion of all millennials who drive a car that is more than 5 years old who have named their cars is between 0.362 and 0.418.

9.29 It is unlikely that the estimated proportion of adult Americans who have pretended to know what the cloud is or how it works ($\hat{p} = 0.22$) will differ from the true proportion by more than 0.026.

9.31 Bigfoot.
Visitors from Other Planets:

$$\text{margin of error} = 1.96\sqrt{\frac{0.39(1-0.39)}{2000}} = 0.0214.$$

Bigfoot:

$$\text{margin of error} = 1.96\sqrt{\frac{0.27(1-0.27)}{2000}} = 0.0195.$$

9.33 A sample size of 2401 is required.

9.35 A sample size of 2017 is required.

9.37 **a.** 2.12 **b.** 1.80 **c.** 2.81 **d.** 1.71 **e.** 1.78 **f.** 2.26

9.39 The second interval is based on the larger sample size; the interval is narrower.

9.41 **a.** If the distribution of volunteer times is approximately normal, for the sample standard deviation of $s = 16.54$ hours and the sample mean of $\bar{x} = 14.76$ hours, approximately 18.6% of volunteer times would be negative, which is impossible. Therefore, it is not reasonable to think that the distribution of volunteer times is approximately normal. **b.** The two conditions required to use a one-sample *t* distribution are that the sample was either randomly selected or is representative of the population, and the population distribution is either normally distributed or that the sample size is at least 30. We are told that the sample is representative of the population, and that the sample size is 500. Both conditions are satisfied in the exercise. **c.** (13.307, 16.213). We are 95% confident that the mean number of hours spent in volunteer activities per year for South Korean middle school children is between 13.307 and 16.213 hours.

9.43 **a.** (35.424, 38.616). We are 95% confident that the mean procrastination scale for first-year students at this college is between 35.424 and 38.616. **b.** Forty (40) is not a plausible value for the mean population procrastination scale score because that value is not contained within the confidence interval. It seems plausible that, on average, students at this university do not have high levels of procrastination.

9.45 **a.** Narrower. **b.** The statement is not correct. The population mean, μ, is a constant, and therefore, we cannot talk about the probability that it falls within a certain interval. **c.** The statement is not correct. We can say that *on average* 95 out of every 100 samples will result in confidence intervals that will contain μ, but we cannot say that in 100 such samples, *exactly* 95 will result in confidence intervals that contain μ.

9.47 **a.** For samples of equal size, the greatest variability in sleep time will be reflected in the wider confidence interval. The confidence interval for the 8:10 A.M. start time is wider (width = 10.4) than the confidence interval for the 7:20 A.M. start time (width = 7.3). Therefore, the 8:10 A.M. start time had the greater variability in sleep time. **b.** For samples with the same standard deviations, larger samples will result in narrower confidence intervals. Therefore, because the width of the confidence interval for the 7:20 A.M. start time (width = 7.3) is smaller than the width of the 8:10 A.M. start time (width = 10.4), the 7:20 A.M. start time sample had the largest sample size. **c.** The width of the interval (462.2, 475.8) is 13.6, which is wider than the confidence interval calculated for the 8:10 A.M. start time sample (463.8, 474.2), which has a width of 10.4. Higher levels of confidence correspond to larger *t* critical values, so the new interval, being wider, must be the 99% confidence interval.

9.49 **a.** (4.386, 4.614) **b.** No, it does not mean that all Generation X Americans have at least one credit card. Rather, it means that we are confident that there are at least some Generation X Americans who have at least one credit card.

9.51 **a.** ($63,061.95, $64,938.05) **b.** Because the confidence interval in Part (a) contains the reported national average of $63,316, we cannot conclude that the mean salary offer for math and science graduates at this university is different from the reported national average.

9.53 A reasonable estimate of σ is given by (sample range)/4 = 162.5. A sample size of 1015 is needed.

9.55 First, we need to know that the information is based on a random sample of Generation X US residents. Second, it would be useful if some sort of margin of error or confidence level were given for the estimated mean of $83,357.

9.57 **a.** The paper states that queens flew for an *average of* 24.2 ± 9.21 minutes on their mating flights, and so this interval is a confidence interval for a population mean. **b.** (3.301, 5.899)

9.59 **a.** No, it is not appropriate to use the large-sample confidence interval for a population proportion to estimate the proportion of the transportation services customers who have tried Uber or Lyft at least once. We don't have at least 10 successes in the sample, which is one of the necessary conditions. **b.** Yes, it is appropriate to use a bootstrap confidence interval for a population proportion to estimate the proportion of the transportation services customers who have tried Uber or Lyft because a random sample of regular customers who are 55 or older was taken. **c.** A bootstrap

95% confidence interval for the population proportion of customers 55 or older who have used Uber or Lyft at least once, p, is (0.000, 0.333). Based on this sample, you can be 95% confident that the actual proportion of customers 55 or older who have used Uber or Lyft at least once is somewhere between 0.000 and 0.333. **d.** Yes, the value obtained in the study (21%) is contained within the bootstrap confidence interval. Confidence intervals provide a range of plausible values for the true proportion. The interval contains 0.21, so we know that 0.21 is a plausible value for the true proportion of customers 55 or older who have used Uber or Lyft at least once.

9.61 **a.** It would not be appropriate to use a large-sample confidence interval for one proportion to estimate the proportion of COVID-19 patients who experience myocardial inflammation because there are only 8 successes in the sample, which is fewer than the required 10 successes to use the large-sample interval. **b.** Different simulations will produce different results, so answers will vary. For one simulation, a bootstrap 90% confidence interval for the population proportion of COVID-19 patients who experience myocardial inflammation, p, was (0.085, 0.255). Based on this sample, you can be 90% confident that the actual proportion of COVID-19 patients who experience myocardial inflammation is somewhere between 0.085 and 0.255.

9.63 **a.** Yes, it would be appropriate to use the large-sample confidence interval for a population proportion to estimate the proportion of residential properties successfully sold at auction in the post-Brexit UK. The necessary conditions (a representative sample of residential properties and at least 10 successes and failures in the sample) have both been satisfied. **b.** It is appropriate to use a bootstrap confidence interval because we have a representative sample of the population. **c.** A bootstrap 95% confidence interval for the population proportion of residential properties successfully sold at auction in the post-Brexit UK, p, is (0.346, 0.731). Based on this sample, you can be 95% confident that the actual proportion of residential properties successfully sold at auction in the post-Brexit UK is somewhere between 0.346 and 0.731. **d.** Yes, the success rate for properties sold at auction throughout the UK during one stretch a year earlier (0.72) is contained within the bootstrap confidence interval.

9.65 Different simulations will produce different results, so answers will vary. For one simulation, a bootstrap 95% confidence interval for the population proportion of all brand-name products that are active on Snapchat, p, is (0.617, 0.783). Based on this sample, you can be 95% confident that the actual proportion of all brand-name products that are active on Snapchat is somewhere between 0.617 and 0.783.

9.67 **a.** A boxplot of the data indicated that the data distribution is skewed so the methods based on the t distribution may not be appropriate. **b.** Different simulations will produce different results, so answers will vary. One simulation produced a confidence interval of (79.827, 202.262). We are 95% confident that the population mean germ score for coins lies somewhere between 79.827 and 202.262.

9.69 The given simulation produced a confidence interval of (4.579, 5.754). We are 95% confident that the population mean annual energy cost for the population of all small televisions lies somewhere between \$4.579 and \$5.754.

9.71 **a.** The sample size of $n = 15$ is smaller than 30, so the methods based on the t distribution may not be appropriate. **b.** Different simulations will produce different results, so answers will vary. One simulation produced a confidence interval of (42, 57). We are 95% confident that the mean interleague winning percentage for National League teams lies somewhere between 42% and 57%. **c.** In Part (b), the mean winning percentage for National League teams was estimated (with 95% confidence) to be between 42% and 57%, which indicates that 50% is a plausible value. Therefore, it is not reasonable to say that the National League or American League performs significantly better than the other in interleague play.

9.73 The meaning of the 90% confidence level refers to the fact that the method used to construct this interval estimate is successful in capturing the actual value of the proportion about 90% of the time. An alternate, but equivalent, interpretation is that 90% of all possible confidence intervals computed from samples of the same size from the population of adult Americans will capture the actual value of the population proportion.

9.75 (0.477, 0.563). We are 90% confident that the true proportion of all homeowners in the Western United States who have considered installing solar panels is between 0.477 and 0.563.

9.77 Assuming a 95% confidence level, and using $p = 0.32$, a sample size of 335 is required. Using the conservative value, $p = 0.5$, a sample size of 385 is required. The conservative estimate of p gives the larger sample size. Since the relevant proportion could have changed significantly since 2011, it would be sensible to use a sample size of 385.

9.79 A sample size of 246 is required.

9.81 (8.571, 9429)

Chapter 10

10.1 $\bar{x}$ is a *sample* statistic.

10.3 H_a: $\mu > 100$ will be used.

10.5 In this case, consider the population of youths with preexisting psychological conditions, and the following hypotheses: H_0: Video games do not increase antisocial behavior *versus* H_a: Video games do increase antisocial behavior. The article states that the researchers found no evidence "that violent video games increase antisocial behavior in youths with pre-existing psychological conditions." Finding no evidence simply means that the data did not provide strong support for the alternative, not that the null hypothesis is true. Therefore, the title is misleading in that it is implying that the null hypothesis is true.

10.7 H_0: $p = 0.5$, H_a: $p > 0.5$

10.9 H_0: $p = 0.5$ H_a: $p > 0.5$

10.11 H_0: $p = 0.7$, H_a: $p \neq 0.7$

10.13 **a.** This is a Type II error because the statement describes the result of failing to reject the null hypothesis when the null hypothesis is actually false (concluding that the woman does not have breast cancer when, in actuality, she does). This probability is approximately $P(\text{Type II error}) \approx \frac{1}{13} = 0.077$. **b.** The other error that is possible is a Type I error, in which a true null hypothesis is rejected. In this scenario, a Type I error is concluding that a woman has cancer when she really does not have cancer. This probability is approximately $P(\text{Type I error}) \approx \frac{90}{637} = 0.141$.

10.15 **a**. A Type I error is when it is concluded that the scat is from a fox when, in fact, it is not from a fox. A Type II error is when it is concluded that the scat is not from a fox when, in fact, it is from a fox. **b**. (b) $\alpha \approx \frac{9}{208} = 0.043$ and $\beta \approx \frac{21}{319} = 0.066$

10.17 **a.** A Type I error concluding that there is evidence that more than 1% of a shipment is defective when in fact (at least) 1% of the shipment is defective. A Type II error is not being convinced that more than 1% of a shipment is defective when in fact more than 1% of the shipment is defective. **b.** Type II **c.** Type I

10.19 The probability of a Type I error is equal to the significance level. Here the aim is to reduce the probability of a Type I error, so a small significance level (such as 0.01) should be used.

10.21 **a.** The phrase "No evidence of increased risk of thyroid cancer in areas that were near a nuclear power plant" indicates that the researchers failed to reject the null hypothesis. **b.** If the researchers are incorrect in their conclusion, then they would have failed to reject H_0 when H_0 was actually false. This is a Type II error. **c.** No. The study did not provide sufficient evidence to reject the null hypothesis. In other words, there was not convincing evidence that the proportion of the population in areas near nuclear power plants who are diagnosed with thyroid cancer during a given year is greater than the proportion of the population in areas where no nuclear power plants are present. This does not mean that there was no effect, only that the data did not provide evidence to reject the null hypothesis.

10.23 **a.** A P-value of 0.0003 means that it is very unlikely (probability = 0.0003), assuming that H_0 is true, that you would get a sample result at least as inconsistent with H_0 as the one obtained in the study. Thus, H_0 is rejected. **b.** A P-value of 0.350 means that it is not particularly unlikely (probability = 0.350), assuming that H_0 is true, that you would get a sample result at least as inconsistent with H_0 as the one obtained in the study. Thus, there is no reason to reject H_0.

10.25 **a.** H_0 is not rejected. **b.** H_0 is not rejected. **c.** H_0 is not rejected.

10.27 **a.** 0.081 **b.** 0.176 **c.** 0.025 **d.** 0.007 **e.** 0.567

10.29 **a.** Based on the title of the article, p represents the proportion of the population of adults from the 116 countries and territories from around the world who would rather live a calm life. **b.** Yes, it is reasonable to say that the data provide strong support for the alternative hypothesis. **c.** Yes, it is reasonable to say that the data provide strong evidence against the null hypothesis.

10.31 **a.** H_0: $p = 0.5$, H_a: $p < 0.5$, $z = -0.376$, P-value = 0.3335, fail to reject H_0. **b.** The same conclusion would be reached if a significance level of 0.01 had been used because the P-value is greater than 0.01, so we would still fail to reject H_0.

10.33 H_0: $p = 0.75$, H_a: $p > 0.75$, $z = 4$, P-value = 0.00003, reject H_0.

10.35 H_0: $p = 0.69$ versus H_a: $p \neq 0.69$, where p = proportion of teens at the high school who access social media from a mobile phone

10.37 H_0: $p = 0.87$, H_a: $p \neq 0.87$

10.39 Yes. H_0: $p = 0.10$, H_a: $p < 0.10$, $z = 2.449$, P-value = 0.007, reject the null hypothesis. There is convincing evidence that the proportion of adult Americans who do not use the Internet is less than 0.10 (10%).

10.41 **a.** The sample must be randomly selected from the population of all British businesses that implemented a 4-day work week. Alternatively, the sample must be representative of the population of all British businesses that implemented a 4-day work week. **b.** H_0: $p = 0.20$, H_a: $p > 0.20$, $z = 1.857$, P-value = 0.0317, reject the null hypothesis. There is convincing evidence to conclude that more than 20% of all British businesses that implemented a 4-day work week would decide to make the change permanent.

10.43 **a.** 0.040 **b.** 0.003 **c.** 0.019 **d.** 0.000

10.45 **a.** 0.484 **b.** 0.686 **c.** 0.025 **d.** 0.000 **e.** 0.097

10.47 **a.** H_0 is rejected. **b.** H_0 is not rejected. **c.** H_0 is not rejected.

10.49 **a.** $t = 0.748$, P-value = 0.468, fail to reject H_0 **b.** $t = 8.731$, P-value ≈ 0, reject H_0

10.51 H_0: $\mu = 10$, H_a: $\mu < 10$, $t = 7.81$, P-value ≈ 0, reject H_0. There is convincing evidence that the mean blood lead level for young children in this county is greater than 10 μg/dL.

Copyright 2025 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s). Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

10.53 H_0: $\mu = 63{,}316$, H_a: $\mu > 63{,}316$, $t = 1.466$, P-value $= 0.075$, fail to reject H_0. There is not convincing evidence that the mean salary offer for math and science graduates of this university is higher than the 2021 national average of $63,316.

10.55 **a.** The fact that the sample standard deviation is greater than the sample mean indicates that there is a quite a bit of variability in the distribution of purchase amounts. In addition, since purchase amounts cannot be negative, the larger standard deviation indicates that the distribution is positively skewed. **b.** H_0: $\mu = 2.80$, H_a: $\mu > 2.80$, $t = 0.857$, P-value $= 0.196$, fail to reject H_0.

10.57 Since the sample was large, it was possible for the hypothesis test to provide convincing evidence that the mean score for the population of children who spent long hours in child care was greater than the mean score for third graders in general, even though the obtained sample mean didn't differ greatly from the known mean for third graders in general.

10.59 **a.** H_0: $\mu = 15$, H_a: $\mu > 15$, $t = 0.999$, P-value $= 0.159$, fail to reject H_0 **b.** H_0: $\mu = 15$, H_a: $\mu > 15$, $t = 1.7315$, P-value $= 0.042$, reject H_0 **c.** The sample standard deviation of 35.5 minutes in Part (a) indicates that there is likely to be more variability in the increase in number of minutes of sleep time than when the sample standard deviation was 20.5 minutes, as in Part (b). If the mean of the population of average increase sleep time was indeed 15 minutes, it is less likely that a sample mean of 17 minutes would be obtained when the standard deviation was 20.5 minutes than when the standard deviation was 35.5 minutes. Therefore, the larger standard deviation in Part (a) makes it more difficult to detect a difference, resulting in a failure to reject the null hypothesis.

10.61 **a.** Yes. Since the pattern in the normal probability plot is roughly linear, and since the sample was a random sample from the population, the t test is appropriate. **b.** The boxplot shows a median of around 245, and since the distribution is a roughly symmetrical distribution, this tells us that the sample mean is also around 245. This might initially suggest that the population mean differs from 240. But, the sample is relatively small, and the sample values range all the way from 225 to 265; such a sample mean would still be feasible if the population mean were 240. **c.** $t = 1.212$, P-value $= 0.251$, fail to reject H_0

10.63 **a.** Increasing the sample size increases the power. **b.** Increasing the significance level increases the power.

10.65 **a.** A Type II error occurs when H_0 is not rejected, and, as shown in the sketch, H_0 will not be rejected for values of $\bar{x}$ less than 152.546. $\beta = 0.374$ **b.** $\beta \approx 0$ **c.** A Type II error could have occurred.

10.67 **a.** P-value $= 0.044$. H_0 is rejected, and we have convincing evidence that more than 75% of apartments exclude children. **b.** 0.351

10.69 **a. i.** $\beta \approx 0.84$; $\beta \approx 0.84$ **ii.** $\beta \approx 0.57$ **iii.** $\beta \approx 0.10$ **iv.** $\beta \approx 0$ **b.** When σ increases, each d value decreases, and, looking at the graphs in Appendix Table 5, we see that each value of β will be greater than in Part (a).

10.71 **a.** $\beta \approx 0$ **b.** $\beta \approx 0.04$ **c.** $\beta \approx 0.01$

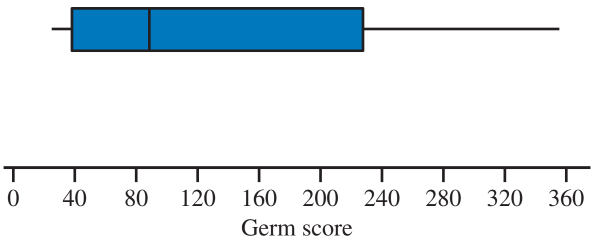
10.73 **a.** Answers will vary. The title of the article is somewhat misleading in that it implies a causal conclusion between eating brownies and lowering triglycerides. I would like to ask the person who wrote the article if the conclusions were based on an experiment or observational study. Additionally, I would like to know if humans and mice respond the same way to eating black beans. **b.** The approximate sample mean is $45(1 - 0.37) = 28.35$. **c.** The main concern is that the sample of mice is quite small ($n = 8$). As such, we would need to know if the distribution of triglyceride levels was approximately normally distributed. If so, then using a one-sample t test is reasonable. **d.** It doesn't seem as if the recommendations in the *Woman's World* article are justified. It seems as if eating a low-fat diet is better than the other two diets.

10.75 **a.** H_0: $p = 0.20$, H_a: $p < 0.20$ **b.** It is not stated that the sample was randomly selected or that the sample is representative of the population of all hospital patients who had been treated for pneumonia using a respiratory therapist protocol. As such, we must use caution and assume the sample was selected in a reasonable way. The large sample size condition has been satisfied because there are at least 10 successes and 10 failures in the sample. **c.** The output for the exact binomial test indicates that the P-value is 0.000. Since the P-value of 0.000 is less than any reasonable significance level, we reject the null hypothesis. We have sufficient evidence to conclude that the proportion of subjects who will be readmitted to a hospital within 30 days after following a respiratory therapist protocol for treatment of pneumonia is less than 0.20. **d.** P-value $= 0.003$, the same conclusion will be drawn.

10.77 **a.** H_0: $p = 0.25$ versus H_a: $p > 0.25$, where p is the proportion of ground chicken packages that contain salmonella. A P-value of 0.145 was obtained. Because the P-value of 0.145 is greater than the significance level of 0.05, we fail to reject the null hypothesis. There is not sufficient evidence to conclude that the proportion of ground chicken packages that contain salmonella is greater than 0.25. **b.** The assumption about the sample of 75 packages is that the sample was randomly selected and therefore representative of the population of all ground chicken packages. **c.** The sample size is large enough to satisfy the large-sample requirement of the large-sample z test for a population proportion. The number of successes (23) is greater than 10, and the number of failures (52) is also greater than 10. Using the large sample test to test the hypotheses in Part (a), a P-value of 0.129 was obtained. We fail to reject the null hypothesis, and there is therefore not sufficient evidence to conclude that the proportion of ground chicken packages that contain salmonella is greater than 0.25. **d.** The same conclusion was reached in Parts (a) and (c). This is not surprising.

10.79 **a.** H_0: $p = 0.50$, H_a: $p > 0.50$, P-value of 0.007. Because the P-value of 0.007 is less than any reasonable significance level, we reject the null hypothesis. We have sufficient evidence to conclude that the proportion of coin flip calls that the Patriots win is greater than 0.50. **b.** In order to carry out the exact binomial version of the hypothesis test, there are four necessary conditions: (1) There is a fixed number of trials; (2) each trial can result in one of only two possible outcomes; (3) outcomes of different trials are independent; and (4) the success probability is the same for each trial. In this case, all the conditions have been satisfied. Rejecting the null hypothesis does not mean that the alternative hypothesis is true. Rather, it means we have evidence that supports the alternative hypothesis based on the data we collected. It could be that we incorrectly rejected the null hypothesis based on this sample (which is a Type I error).

10.81 **a.** A boxplot of the data is positively skewed, so the methods based on the *t* distribution might not be appropriate.



b. (b) H_0: $\mu = 50$ versus H_a: $\mu > 50$. Different simulations will produce different results, so answers will vary. One randomization test resulted in a P-value of 0.013. Because this P-value of 0.013 is less than a significance level of 0.05, we reject the null hypothesis. We have convincing evidence that the population mean germ score for coins is greater than 50.

10.83 **a.** The given randomization test resulted in a P-value of 0.000. Because this P-value of 0.000 is less than a significance level of 0.05, we reject the null hypothesis. We have convincing evidence that the population mean 2000-meter ergometer time for U.S. junior male sculls rowers differs from the 2007 international standard of 387 seconds. **b.** Based on the result of the hypothesis test, we can say that there is evidence to conclude that the mean 2000-meter ergometer time for U.S. junior male sculls rowers differs from the 2007 international standard of 387 seconds. The sample mean time of 394.25 seconds, given in the Shiny App output, is greater than 387 seconds. Therefore, since the null hypothesis was rejected and the sample mean is greater than the hypothesized mean, we can conclude that the U.S. junior male sculls rowers seem to have "caught up," on average, with the international championship rowers from 2007.

10.85 **a.** The fact that the sample size ($n = 15$) is smaller than 30 indicates that the methods based on the *t* distribution might not be appropriate. **b.** Different simulations will produce different results, so answers will vary. One randomization test resulted in a P-value of 0.82. Because this P-value of 0.82 is greater than a 5% significance level, we fail to reject the null hypothesis. We do not have convincing evidence that the population mean interleague winning percentage for National League teams differs from 50%. **c.** In Part (b), the null hypothesis was not rejected, which indicates that there is insufficient evidence to conclude that the mean winning percentage for National League teams differs from 50%. Therefore, it is not reasonable to say that the National League or American League performs significantly better than the other in interleague play.

10.87 **a.** (0.498, 0.542). We are 95% confident that the proportion of adult Americans who grew up sleeping with a security blanket or stuffed animal is between 0.498 and 0.542. **b.** The confidence interval from Part (a) does include 0.5. Therefore, it is not reasonable to conclude that a majority of adult Americans grew up sleeping with a security blanket or stuffed animal.

10.89 H_0: $\mu = 7.85$, H_a: $\mu > 7.85$, $t = 3.28$, P-value $= 0.0013$, reject H_0.

10.91 **a.** We would need to assume that the sample was randomly selected from the population of white-collar workers in the United States or that the sample is representative of the population. **b.** H_0: $\mu = 4$, H_a: $\mu > 4$, $t = 2.437$, P-value $= 0.0075$, reject H_0.

10.93 (0.660, 0.720). We are 99% confident that the proportion of adult Americans who would be willing to wear a fitness tracker if it would give them a discount on their health insurance is between 0.660 and 0.720.

10.95 $z = 1.069$, P-value $= 0.143$, fail to reject H_0

10.97 $z = -4.032$, P-value $= 0.00003$, reject H_0

10.99 $t = -5.324$, P-value ≈ 0, reject H_0

Cumulative Review Exercises 10

CR10.3 **a.** Two airlines stand out from the rest and have large numbers of canceled flights. These airlines are United with 3566 canceled flights and Delta with 2988 canceled flights. **b.** A typical number of canceled flights per 1000 flights is around 32, with most lying between 11.3 and 62.8. Four airlines stand out from the rest and have high rates. **c.** The rate per 1000 flights data should be used, since this measures the likelihood of any given flight being late. An airline could stand out in the number of flights delayed data purely as a result of having a large number of flights.

CR10.5 **a.** 0.148 **b.** 0.032 **c.** $\mu_x = 22$, $\sigma_x = 4.271$ **d.** 0.146

CR10.7 **a.** 0.22 **b.** 0.12 **c.** 0.65 **d.** 0.078.

CR10.9 **a.** H_0: $p = 0.5$ versus H_a: $p > 0.5$, where p is the proportion of adult Americans who are confident in their ability to recognize fake news. **b.** Reject the null hypothesis **c.** I disagree with the friend in my statistics class.

Without taking a random sample, regardless of the P-value, we worry that the sample could suffer from some sort of bias. For example, perhaps this was a voluntary response sample, and only those who feel strongly that they can recognize fake news responded. Those who don't feel that they can recognize fake news are not included, leading to a misleading conclusion.

CR10.11 (0.831, 0.869). We are 90% confident that the proportion of all adult Americans who believe that civility is worse compared to 10 years ago is between 0.831 and 0.869.

CR10.13 $z = 2.45$, P-value $= 0.007$, reject H_0

CR10.15 The formula for standard error ($s/\sqrt{n}$) has sample size in the denominator. Notice, also, that the sample size for normal birth weight babies ($n = 73$) is larger than the sample size for low birth weight babies ($n = 50$). Larger values in the denominator result in smaller values for the standard error, which would explain the difference in the two standard errors.

Chapter 11

11.1 The distribution of $\bar{x}_1 - \bar{x}_2$ is approximately normal with mean 5 and standard deviation 0.529.

11.3 H_0: $\mu_1 - \mu_2 = 0$, H_a: $\mu_1 - \mu_2 \neq 0$, $t = -4.536$, df $= 71$, P-value ≈ 0, reject H_0. There is convincing evidence that the mean reading level for health-related pages differs for Wikipedia and WebMD.

11.5 **a.** The large standard deviation tells us that there is a large amount of variability in the amount of money allocated to stock investments around the respective group means among both dog owners and cat owners. **b.** $H_0 : \mu_D - \mu_C = 0$, $H_a : \mu_D - \mu_C > 0$, $t = 2.32$, P-value $= 0.011$, reject the null hypothesis. There is convincing evidence that the mean amount of money allocated to stack investments is greater for dog owners than cat owners. **c.** Yes, the hypothesis that dog owners would exhibit a greater tolerance for risk is supported by the result of the hypothesis test in Part (b) because the null hypothesis of no difference between dog owners and cat owners was rejected in favor of the alternative hypothesis, which states that the mean amount of money allocated to stack investments is greater for dog owners than cat owners.

11.7 $H_0 : \mu_M - \mu_H = 0$, $H_a : \mu_M - \mu_H \neq 0$, $t = 0.7896$, P-value $= 0.433$, fail to reject the null hypothesis. There is not convincing evidence that the mean amount of time spent on Facebook is different for older adults with medium mobility and the mean for older adults with high mobility.

11.9 $H_0 : \mu_1 - \mu_2 = 0$, $H_a : \mu_1 - \mu_2 \neq 0$, $t = 1.965$, P-value $= 0.054$, fail to reject the null hypothesis. There is not convincing evidence that the mean blood lead concentration is different for Atlantic Coast golden eagles and Pacific Coast golden eagles.

11.11 **a.** If the distributions were both approximately normal, then approximately 16% of the germ scores for credit cards would be negative, and approximately 15% of the germ scores for bills and coins would be negative. Assuming that the scores must be positive, these percentages are too large to make it reasonable to believe that the distributions are approximately normal. **b.** It would be appropriate to use the two-sample t test to test the null hypothesis that there is no difference in the mean germ scores for credit cards and money. The sample sizes are both larger than 30, so the sampling distribution of the difference in sample means is approximately normal, which allows us to carry out the test. **c.** $H_0 : \mu_C - \mu_M = 0$, $H_a : \mu_C - \mu_M > 0$, $t = 2.617$, P-value $= 0.0056$, reject the null hypothesis. There is convincing evidence that the mean germ score is greater for credit cards than for money.

11.13 $(-1.671, -0.704)$. We can be 95% confident that the difference in the mean sit-stand-sit time for the population of stroke patients who receive biofeedback weight training for 8 weeks and the population of stroke patients who did not receive biofeedback weight training for 8 weeks is between -1.671 and -0.704 seconds. Both endpoints of this interval are negative, so you can estimate that the mean sit-stand-sit time for the population of stroke patients who receive biofeedback weight training for 8 weeks is less than the mean sit-stand-sit time for the population of stroke patients who did not receive biofeedback weight training for 8 weeks by somewhere between 0.7035 and 1.6705 seconds.

11.15 **a.** We would need to assume that the samples were independently selected and representative of their respective populations. **b.** $H_0 : \mu_1 - \mu_2 = 0$, $H_a : \mu_1 - \mu_2 > 0$, $t = 7.130$, P-value ≈ 0, reject the null hypothesis. There is convincing evidence that the mean length of sleep on a school night is greater for students with a late school start time than for students with an early school start time. **c.** Allowing participants to self-select whether they wanted to be in the late start time sample or the early start time sample could introduce systematic differences between the two groups. Random assignment ensures that a study does not favor one condition (treatment) over another and tries to create groups that are as much alike as possible.

11.17 The difference in sample mean scores on the measure of working memory for the two groups could have occurred by chance just due to the random assignment.

11.19 **a.** Since the samples are small it is necessary to know—or to assume—that the distributions from which the random samples were taken are normal. However, in this case, since both standard deviations are large compared to the means, it seems unlikely that these distributions would have been normal. **b.** Since the samples are large, it is appropriate to carry out the two-sample t test. **c.** $t = -2.207$, P-value $= 0.030$, fail to reject H_0

11.21 **a.** $t = -9.863$, P-value ≈ 0, reject H_0 **b.** For the two-sample t test, $t = -9.979$, $df = 22.566$, and P-value ≈ 0. Thus, the conclusion is the same.

11.23 H_0: $\mu_d = 0$, H_a: $\mu_d > 0$. $t = -0.247$, P-value $= 0.595$, fail to reject H_0. There is not convincing evidence that the mean time to exhaustion for experienced triathletes is greater when they run while listening to motivational music.

11.25 $(-58.198, 46.598)$. We can be 95% confident that the difference in mean time to exhaustion for experienced triathletes when running to motivational music and the mean time when running with no music is somewhere between -58.198 and 46.598 seconds. Because the endpoints of the confidence interval have opposite signs, zero is included in the interval, and there may be no difference in the mean time to exhaustion for experienced triathletes when running to motivational music and the mean time when running with no music.

11.27 H_0: $\mu_d = 0$, H_a: $\mu_d > 0$. $t = 2.80$, P-value $= 0.005$, reject H_0. There is convincing evidence that the mean bleeding score is lower after 2 months of kiwi consumption.

11.29 H_0: $\mu_d = 0$, H_a: $\mu_d > 0$. $t = 2.724$, P-value $= 0.004$, reject H_0. There is convincing evidence to support the claim that the mean following distance for Greek taxi drivers is greater when there are no distractions than when the driver is talking on a mobile phone. This conclusion is consistent with the claim made by the authors.

11.31 (0.862, 1.738). We can be 95% confident that the mean following distance for Greek taxi drivers while driving with no distractions and while driving and texting is between 0.862 and 1.738 meters. Both endpoints of this interval are positive, so you can estimate that the mean following distance for Greek taxi drivers while driving with no distractions and while driving and texting is somewhere between 0.862 and 1.738 meters.

11.33 $t = -0.515$, P-value $= 0.612$, fail to reject H_0

11.35 $t = 0.639$, P-value $= 0.267$, fail to reject H_0

11.37 $t = 0.225$, P-value $= 0.825$, fail to reject the null hypothesis. There is not convincing evidence that the mean reading differs for the Mini-Wright meter and the Wright meter.

11.39 **a.** $t = 4.321$, P-value ≈ 0, reject H_0 **b.** $t = 1.662$, P-value $= 0.055$, fail to reject H_0 **c.** A smaller standard deviation in the sample of differences means that we have a lower estimate of the standard deviation of the population of differences. Assuming that the mean wrist extensions for the two mouse types are the same (in other words, that the mean of the population of differences is zero), a sample mean difference of as much as 8.82 is much less likely when the standard deviation of the population of differences is around 10 than when the standard deviation of the population of differences is around 26.

11.41 $z = -3.086$, P-value $= 0.001$, reject H_0. There is convincing evidence that the proportion of people for whom the wart is successfully removed is less for the liquid nitrogen treatment than for the duct tape treatment.

11.43 $z = 0.624$, P-value $= 0.533$, fail to reject H_0. We do not have convincing evidence that the proportion who require students to submit papers through plagiarism-detection software is different for full-time faculty and part-time faculty.

11.45 No, I would not use the large-sample test for a difference in population proportions because this question is phrased as a one-proportion hypothesis test, not a difference in proportions.

11.47 (0.033, 0.081). We are 90% confident that the actual difference in proportions living in poverty in 2020 for these two age groups is somewhere between 0.033 and 0.081. Because both endpoints of the confidence interval are positive, we believe that the percent of people in the under age 18 group living in poverty is greater than the percent of people in the 18 to 64 age group living in poverty by somewhere between 3.3% and 8.1%.

11.49 **a.** (0.175, 0.425). We are 95% confident that the difference in the proportions of people who would report that their neck pain improved for those who are given the app-based relaxation exercises and those who are not is somewhere between 0.175 and 0.425. Both endpoints of the confidence interval are positive, so you can estimate that the proportion of people who would report that their neck pain improved using the app-based relaxation exercises is greater than the proportion of people who would report that their neck pain improved who do not use the app-based relaxation exercises by somewhere between 0.175 and 0.425. **b.** Yes, the interval in Part (a) does support the conclusion that people using app-based relaxation exercises are more likely to report improvement. The confidence interval for the difference in proportions (app-based relaxation exercises minus no app-based relaxation exercises) for those who would report improvement has endpoints that are positive, indicating that the app-based relaxation exercises group tends to report higher levels of neck pain improvement than the no app-based relaxation exercises group.

11.51 $z = -0.926$, P-value $= 0.354$, fail to reject the null hypothesis. There is not convincing evidence that the proportion of people who think that, in general, scientists in their country are trustworthy differs for the United States and Canada.

11.53 **a.** Yes, sample size conditions are met. **b.** (0.027, 0.173). We are 90% confident that the actual difference in the proportion of adult Americans age 18 to 29 who believe the foods made with genetically modified ingredients are bad for their health and the corresponding proportion for adult Americans age 50 to 64 is somewhere between 0.027 and 0.173. **c.** Zero is not included in the confidence interval. Because both endpoints of the confidence interval are positive, we believe that the percent of adult Americans age 18 to 29 who believe the foods made with genetically modified ingredients are bad for their health is greater than the corresponding proportion for adult Americans age 50 to 64 by somewhere between 2.7 and 17.3 percentage points.

11.55 $z = 1.54$, P-value $= 0.12$, fail to reject the null hypothesis. There is not convincing evidence that the proportion who experience choice blindness is different for the two treatments (choice based on a picture and choice based on seeing the actual candy).

11.57 **a.** $(-0.0536, 0.0256)$. We are 99% confident that the actual difference in the proportion of people with a Master's degree who were unemployed in 2020 and people with a Bachelor's degree who were unemployed in 2020 is somewhere between -0.0536 and 0.0256. Because the endpoints of the confidence interval have opposite signs, zero is included in the interval, and there may be no difference in the proportion of people with a Master's degree who were unemployed in 2020 and people with a Bachelor's degree who were unemployed in 2020. **b.** Wider, because the confidence level in Part (a) is greater than the confidence level in the previous exercise (a higher confidence level results in a larger z critical value multiplier), and one of the sample sizes is smaller.

11.59 H_0: $p_1 - p_2 = 0$, H_a: $p_1 - p_2 > 0$, $z = 3.76$, P-value ≈ 0, reject H_0. There is convincing evidence that the proportion of Gen Xers who do not pay off their credit cards each month is greater than this proportion for millennials.

11.61 **b.** If we want to know whether the e-mail intervention *reduces* (as opposed to *changes*) adolescents' display of risk behavior in their profiles, then we use one-sided alternative hypotheses and the P-values are halved. If that is the case, using a 0.05 significance level, we are convinced that the intervention is effective with regard to reduction of references to sex and that the proportion showing any of the three protective changes is greater for those receiving the e-mail intervention. Each of the other two apparently reduced proportions could have occurred by chance.

11.63 Different simulations will produce different results, so answers will vary. For one randomization test, results indicated a P-value of 0.001. Because this P-value of 0.001 is less than the significance level of 0.05, the null hypothesis is rejected. There is convincing evidence that the difference in the population mean heart rate percentage for drivers in a two-firefighter team is greater than the mean heart rate for drivers in a team with more than two firefighters

11.65 Different simulations will produce different results, so answers will vary. For one randomization test, results indicated a P-value of 0.001. Because this P-value of 0.001 is less than a significance level of 0.05 or 0.01, the null hypothesis is rejected at either of these significance levels. There is convincing evidence that the mean Personal Meaning Score for patients taking a high dose is greater than the mean score for patients taking a low dose.

11.67 **a.** Because the sample sizes for each treatment group are smaller than 30, we must be wary of using the two-sample t methods. Boxplots for ADHD Severity Scores show that the data are skewed. **b.** Different simulations will produce different results, so answers will vary. For one randomization test, results indicated a P-value of 0.000. Because this P-value of 0.000 is less than a significance level of 0.05, the null hypothesis is rejected. There is convincing evidence that the mean 1-year improvement in ADHD Severity Score for the ONTRAC treatment is different than the mean 1-year improvement in ADHD Severity Score for the control treatment.

11.69 **a.** The sample size is not greater than 30 and a dotplot of the difference data suggests that it may not be reasonable to assume that the difference distribution is normal. **b.** Different simulations will produce different results, so answers will vary. For one randomization test, results indicated a P-value of 0.003. Because this P-value of 0.003 is less than any reasonable significance level, the null hypothesis is rejected. There is convincing evidence that the mean difference in movement is greater than 0.

11.71 **a.** In order to use the large-sample hypothesis test for the difference in two population proportions, the number of successes and failures in each sample must be at least 10. In this case, the number of successes in one of the groups is $n_1\hat{p}_1 = (21)(0.05) = 1.05$, which is less than 10. **b.** For the given simulation, results indicate a P-value of 0.002. Because this P-value of 0.002 is less than any reasonable significance level, we reject the null hypothesis. We have convincing evidence that the proportion of Tasmanian devils with the genetic marker was greater after DFTD than before DFTD. **c.** The given simulation produced a confidence interval of $(-0.548, -0.214)$. We are 95% confident that the true difference in the rates of occurrence of the specific genetic marker in the genes of Tasmanian devils, before and after DFTD, lies somewhere between -0.548 and -0.214.

11.73 **a.** Data from this study should not be analyzed using a large-sample hypothesis test for a difference in two population proportions because the number of successes in each group (5 in the high-intensity group and 0 in the regular-intensity group) are both less than 10. **b.** Different simulations will produce different results, so answers will vary. For one randomization test, results indicated a P-value of 0.017. Because this P-value of 0.017 is less than any reasonable significance level, we reject the null hypothesis. We have convincing evidence of a difference in the population proportions of patients in the two exercise groups (high-intensity and regular-intensity) who die within 1.5 years. **c.** Different simulations will produce different results, so answers will vary. One simulation resulted in a confidence interval of (0.123, 0.133). We are 95% confident that the actual difference in the population proportions of patients who die within 1.5 years for the two exercise groups is somewhere between 0.123 and 0.133.

11.75 **a.** The given randomization test resulted in a *P*-value of 0.000. Because this *P*-value of 0.000 is less than any reasonable significance level, we reject the null hypothesis. We have convincing evidence that the proportion who think it is OK to use a cell phone at a restaurant is higher for the 18 to 29 age group than for the 30 to 49 age group. **b.** The given simulation resulted in a confidence interval of (0.058, 0.139). We are 90% confident that the actual difference in the proportions of people age 18 to 29 and those age 30 to 49 who would say that it is acceptable to use a cell phone in a restaurant is somewhere between 0.058 and 0.139. **c.** No, the interpretation would not change. The endpoints of the confidence interval change slightly—from Example 11.1, the confidence interval is (0.061, 0.139)—but the conclusion and interpretation remain the same. **d.** (0.060, 0.140). We are 90% confident that the actual difference in the proportion of people who think it is okay to use a cell phone in a restaurant for 18- to 29-year-olds and for 30- to 49-year-olds is between 0.060 and 0.140. This means that we can be confident that the proportion is greater for 18- to 29-year-olds than for 30- to 49-year-olds by somewhere between 0.060 and 0.140. The interpretation would not change. The endpoints of the confidence interval change slightly, but the conclusion and interpretation remain the same.

11.77 **a.** (0.091, 0.265). We are 90% confident that the difference in the population proportions of cell phone users age 20 to 39 and those age 40 to 49 who say that they sleep with their cell phones is between 0.091 and 0.265. **b.** The given simulation resulted in a bootstrap confidence interval of (0.093, 0.267). Therefore, we are 90% confident that the difference in the population proportions of cell phone users age 20 to 39 and those age 40 to 49 who say that they sleep with their cell phones is between 0.093 and 0.267. **c.** The confidence intervals in Parts (a) and (b) are very similar. The lower endpoints are 0.091 and 0.093, respectively, and the upper endpoints are 0.265 and 0.267, respectively. The interpretations are essentially the same, with the only differences due to the slight differences in the endpoints.

11.79 We are 95% confident that the difference in mean scores (neurosurgeons – aerospace engineers) on a spatial problem-solving task lies between −0.39 and 0.01. Because the endpoints of the confidence interval have opposite signs, zero is included in the interval, and there may be no difference in the scores (neurosurgeons – aerospace engineers) on the spatial problem-solving task.

11.81 **a.** $z = 1.011$, *P*-value $= 0.312$, fail to reject the null hypothesis. There is not convincing evidence that there is a difference between the proportion of Boomers and the proportion of Generation X who chose browsing the Internet as their favorite entertainment activity. **b.** $z = 2.021$, *P*-value $= 0.043$, reject the null hypothesis. There is convincing evidence that there is a difference between the proportion of Boomers and the proportion of Generation X who chose browsing the Internet as their favorite entertainment activity. **c.** Assuming that the population proportions are equal, you are much less likely to get a difference in sample proportions as large as the one given when the samples are very large than when the samples are relatively small.

11.83 **a.** Two-sample t test **b.** $t = -8.01$, *P*-value ≈ 0, reject the null hypothesis. There is convincing evidence that the mean attractiveness rating is not the same for the two experimental groups. **c.** The conclusion of the paper is consistent with the results of the test from Part (b). In Part (b), we rejected the null hypothesis of the faces being equally attractive. We had sufficient evidence to conclude that there was a difference in attractiveness between the two groups (no mask and medical mask groups).

11.85 When the null hypothesis is not rejected, it means that you don't have convincing evidence that the treatment proportions are different.

11.87 **a.** (0.134, 0.226). We are 95% confident that the actual difference in proportion of teens who think they spend too much time on their cell phones and this proportion for parents of teens is somewhere between 0.134 and 0.226. Because both endpoints of the confidence interval are positive, we believe that the proportion of teens who think they spend too much time on their cell phones is greater than this proportion for parents of teens by somewhere between 13.4% and 22.6% **b.** No, zero is not included in the confidence interval. This implies that there may be a difference in the proportion of teens who think they spend too much time on their cell phones and this proportion for parents of teens.

11.89 **a.** (−4.738, 22.738). **b.** $t = 0.140$, *P*-value $=$ 0.890, fail to reject H_0 **c.** $t = -0.446$, *P*-value $= 0.330$, fail to reject H_0

11.91 **a.** (−0.186, 0.111). We are 90% confident that the mean difference between surface and subsoil pH is between −0.186 and 0.111. **b.** We must assume that the distribution of differences across all locations is normal.

Chapter 12

12.1 **a.** Fail to reject H_0 **b.** Reject H_0 **c.** Reject H_0

12.3 $X^2 = 133.233$, *P*-value ≈ 0, reject H_0. There is convincing evidence that one or more of the age groups buys a disproportionate share of Texas Lottery tickets.

12.5 **a.** $X^2 = 79.6$, *P*-value ≈ 0, reject H_0. There is convincing evidence that these three work modes are not equally preferred. **b.** For those workers who prefer entirely in-person and hybrid-remote (partially remote), the observed counts are significantly higher than the numbers that would have been expected if the work modes were equally preferred, while for entirely remote, the observed count is much lower than the number that would have been expected.

12.7 $X^2 = 236.356$, P-value ≈ 0, reject H_0. There is convincing evidence that the response proportions are not all the same for the population of millennials.

12.9 $X^2 = 11.36$, P-value $= 0.0099$, reject H_0. We have convincing evidence that the proportions of form of payment used most often differs from the claimed proportions.

12.11 Since the P-value is small, there is convincing evidence that the population proportions of people who respond "Monday" and "Friday" are not equal.

12.13 **a.** $X^2 = 4.63$, P-value $= 0.099$, fail to reject H_0. We do not have convincing evidence that the theory is incorrect. **b.** The analysis and conclusion would be the same.

12.15 **a.** P-value $= 0.844$, fail to reject H_0 **b.** P-value $= 0.106$, fail to reject H_0

12.17 Answers may vary; one example follows. A political science researcher might be interested in seeing if there is evidence that the distribution of self-described political philosophy (liberal, moderate, conservative) among registered voters has changed over the years 2008, 2009, 2010, and 2011. Therefore, the populations that would be sampled are registered voters in each of the years 2008, 2009, 2010, and 2011. The variable that would be recorded is political philosophy (liberal, moderate, or conservative).

12.19 **a.** $X^2 = 13.814$, P-value $= 0.001$, reject H_0. There is convincing evidence that the variables tipping response and income are not independent. **b.** The particularly high contributions to the chi-square statistic come from the "Do Not Always Tip" response. Within this response, there was a disproportionately high number of those who do not always tip in the low-income category and a disproportionately low number of those who do not always tip in the high-income category.

12.21 $X^2 = 511.09$, P-value ≈ 0, reject H_0. There is convincing evidence that there is an association between age group and frequency of illegal smartphone use while driving.

12.23 $X^2 = 29.211$, P-value ≈ 0, reject H_0. There is convincing evidence that the trustworthiness response distribution is not the same for the real face and the synthesized face.

12.25 **a.** $X^2 = 29.757$, P-value ≈ 0, reject H_0. There is convincing evidence that the proportions who would fall into each of the two response categories (did experience quarter-life crisis and did not experience quarter-life crisis) are not the same for Generation Z and millennials. **b.** Yes, the calculations and conclusions from Part (a) are consistent with the Minitab output. The expected counts, X^2 statistic and P-value are the same as what is shown in Part (a). Because the P-value is the same as in Part (a), we would again reject the null hypothesis. **c.** The two-sample z test leads to the same conclusion. In this case, the P-value of approximately 0 is less than the significance level of $\alpha = 0.05$, so we reject the null hypothesis. There is sufficient evidence to conclude that the proportions who would fall into each of the two response categories (did experience quarter-life crisis and did not experience quarter-life crisis) are not the same for Generation Z and millennials. **d.** The P-values in Parts (a) and (c) are exactly the same. This is not surprising because the analysis can be done using either the test of homogeneity or the two-sample z test, and the same data should yield the same conclusion.

12.27 $X^2 = 190.99$, P-value ≈ 0, reject H_0. There is convincing evidence that the race/ethnicity proportions differ for students and teachers.

12.29 **a.** H_0: Age and fruit consumption are independent H_a: Age and fruit consumption are not independent **b.** P-value ≈ 0, reject H_0. We have convincing evidence that age and fruit consumption are not independent. **c.** The same conclusion would have been reached because the P-value is essentially zero, which is less than 0.01. **d.** $X^2 = 59.324$ P-value ≈ 0, reject H_0. We have convincing evidence that age and vegetable consumption are not independent. The differences between the observed and expected counts are consistent with the statement in the paper that says that 1-year-old children were more likely to eat vegetables daily than older children. The positive differences occur for 1- and 2-year-old children who eat vegetables daily, and the negative differences occur for older children who eat vegetables daily.

12.31 $X^2 = 9.2344$, P-value $= 0.06$, fail to reject H_0. There is not convincing evidence that the distribution of responses to the question is not the same for cat owners and dog owners.

12.33 $X^2 = 45.533$, P-value ≈ 0, reject H_0. There is convincing evidence that there is an association between political philosophy and the response to the question about whether a dog or a cat would make a better president.

12.35 $X^2 = 98.526$, P-value $= 0.000$, reject H_0. There is convincing evidence that there is an association between age and whether or not news about privacy-related issues is followed.

Chapter 13

13.1 **a.** Deterministic **b.** Probabilistic **c.** Probabilistic **d.** Deterministic

13.3 **a.** 30.7 therms **b.** 0.017 therms **c.** 1.7 therms **d.** No, 500 sq. ft. is outside the range of x values in the data set.

13.5 **a.** For each one-unit increase in horsepower the predicted fuel efficiency decreases by 0.150 mpg. **b.** The number $\hat{y} = 29.0$ is both an estimate of the mean fuel efficiency when the horsepower is 100 and a prediction for the horsepower of a car whose horsepower is 100. **c.** When $x = 300$, $\hat{y} = -1.0$. This result cannot be valid, since it is not possible to have a car whose fuel efficiency is negative. This has probably occurred because 300 is outside the range of horsepower ratings for the small cars used in the sample; the estimated regression equation is not valid for values outside this range.

Copyright 2025 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s). Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

13.7 **a.** 47, 4700 **b.** 0.3156, 0.0643

13.9 **a.** $r^2 = 23.48\%$ **b.** The point estimate of σ_e is $s_e = 5.36559$. The typical amount by which the depression score change value deviates from the value in the sample and what was predicted using the least-squares regression line is 5.366. **c.** 5.08 **d.** 12.966

13.11 **a.** The coefficient of determination is $r^2 = 0.0204$. Approximately 2.04% of the observed variability in Overall Score can be explained by (or attributed to) the model relationship between Price per Ounce and Overall Score. **b.** The low r^2 value (2.04%) is consistent with what is seen in the scatterplot in that there is not a strong linear relationship between the two variables. In fact, the relationship between Overall Score and Price per Ounce is very weak. Notice as well that the *P*-value (0.413) testing if the slope of the least-squares regression line is significantly different from zero is greater than any reasonable significance level, which indicates that we don't have evidence of a non-zero slope of the least-squares regression line.

13.13 **a.** The plot shows a linear pattern, and the vertical spread of points does not appear to be changing over the range of *x* values in the sample. If we assume that the distribution of errors at any given *x* value is approximately normal, then the simple linear regression model seems appropriate. **b.** $\hat{y} = -0.00227 + 1.247x$; when $x = 0.09$, $\hat{y} = 0.110$. **c.** $r^2 = 0.436$, 43.6% of the variation in market share can be explained by the linear regression model relating market share and advertising share.
d. $s_e = 0.0263$, df $= 8$

13.15 **a.** Yes. *P*-value $= 0.023$. Since this value is less than 0.05, we have convincing evidence at the 0.05 significance level of a useful linear relationship between shrimp catch per tow and oxygen concentration density.
b. Yes. $r = \sqrt{0.496} = 0.704$. So there is a moderate to strong linear relationship for the values in the sample.
c. (17.363, 177.077) We are 95% confident that the slope of the population regression line relating mean catch per tow to O_2 saturation is between 17.363 and 177.077.

13.17 **a.** 0.1537 **b.** (2.17, 2.83) **c.** Yes, the interval is relatively narrow.

13.19 **a.** $a = 592.1$, $b = 97.26$ **b.** When $x = 2$, $\hat{y} = 786.62$, $y - \hat{y} = -29.62$. **c.** (87.76, 106.76)

13.21 $t = -3.66$, *P*-value ≈ 0, reject H_0

13.23 **a.** (0.081, 0.199) We are 95% confident that the mean change in pleasantness rating associated with an increase of 1 impulse per second in firing frequency is between 0.081 and 0.199. **b.** $t = 5.451$, *P*-value $= 0.001$, reject H_0

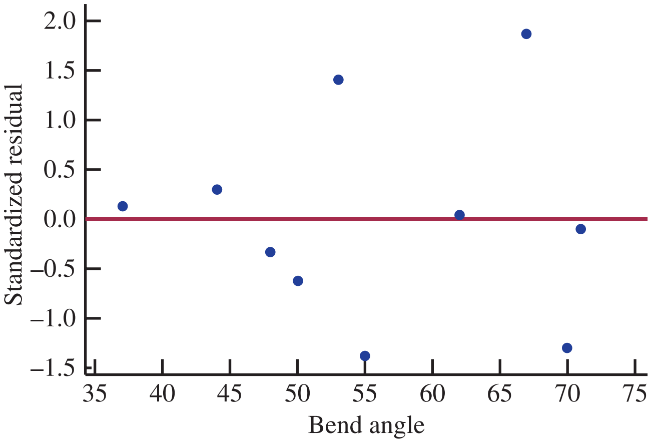
13.25 **a.** This is the slope of the estimated regression line, or $b = 0.359$ μg/kg. **b.** When $x = 250$ s, the predicted acrylamide concentration is 176.75 μg/kg. **c.** $t = 0.82$, *P*-value $= 0.459$, fail to reject H_0

13.27 **a.** $t = 6.493$, *P*-value ≈ 0, reject H_0 **b.** $t = 1.56$, *P*-value $= 0.079$, fail to reject H_0

13.29 **a.** 0.253 **b.** 0.179; no **c.** 4

13.31 Yes. The pattern in the scatterplot looks linear and the spread around the line does not appear to be changing with *x*. There is no obvious pattern in the residual plot and the boxplot of the residuals is approximately symmetric and there are no outliers.

13.33 **a.**



There do not appear to be any unusually large residuals. Additionally, there do not seem to be any influential observations.
b. The only pattern observed in the residual plot is that the spread of the standardized residuals tends to increase as bend angle increases. As such, the simple linear regression model might not be appropriate.

13.35 If the point (69.99, 184.95) is not included, then the slope of the least-squares line would be less than when the point is included (in other words, a shallower but still positive slope).

Cumulative Review Exercises 13

CR13.1 **a.** Observational study because the children were not assigned at random to a treatment group (i.e., a particular diet). **b.** It is not reasonable to conclude that adherence to a MIND diet is the cause of better performance because this was an observational study and not an experiment.

CR13.3 **a.** When comparing 2018 to 2023, there has been an overall decrease in the percentage who responded all of the time or most of the time to the question about trusting the Supreme Court to do the right thing from 44% to 33%. **b.** There are decreases in each of the political affiliations. However, the observed decreases are not the same for each group. The biggest decrease in the percentage who responded all of the time or most of the time to the question about trusting the Supreme Court to do the right thing occurred among those respondents who identify as Democrats (15% decrease), followed by Independents (11% decrease), and then followed by Republicans (3% decrease).

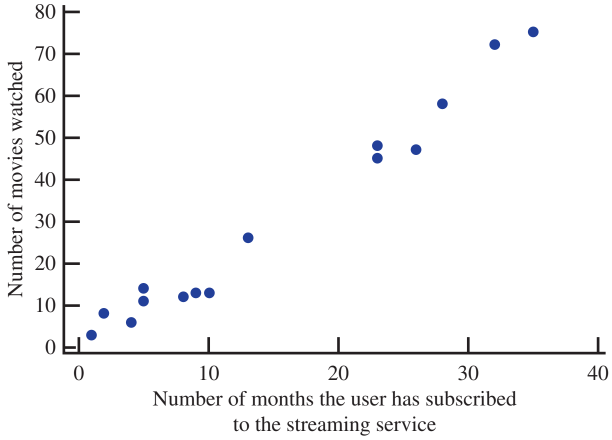
CR13.5 **a.** I would want to know if the sample was randomly selected and from what population the sample was taken. **b.** $z = -1.386$, *P*-value $= 0.0829$, fail to reject H_0.

Copyright 2025 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s). Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

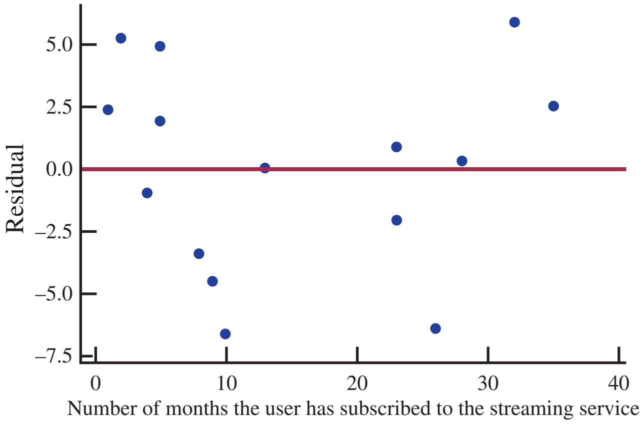
We do not have convincing evidence that the proportion of young adults that are living at home with their parents is less than 0.50.

CR13.7 No, the oval shape distorts the segments and the area is no longer proportional to the percentages being represented.

CR13.9 **a.**



Yes, the relationship looks approximately linear.
b. $\hat{y} = -1.477 + 2.1123x$, where $x =$ Number of months the user has subscribed to the streaming service and $y =$ Number of movies watched on a streaming service.
c. There is a random pattern in the residual plot and there is no suggestion that the variance of y is not the same at each x value. There are no outliers. The assumptions of the simple linear regression model would therefore seem to be reasonable.



d. $t = 22.1$, P-value ≈ 0, reject H_0. We have convincing evidence of a useful linear relationship between the number of movies watched on a streaming service and the number of months the user has subscribed to the streaming service.

CR13.11 **a.** $X^2 = 16.66$, P-value ≈ 0, reject H_0. We have convincing evidence that the response distributions are not the same for 2011 and 2022. **b.** Yes, the conclusion is consistent with the results of the hypothesis test in Part (a). We have convincing evidence that the response distributions are not the same for 2011 and 2022. Additionally, the number of respondents who answered "Yes" has decreased, the number of respondents who answered "No" has increased, and the number who replied "Unsure" has decreased.

CR13.13 $X^2 = 4.8$, P-value $= 0.684$, fail to reject H_0

Chapter 14

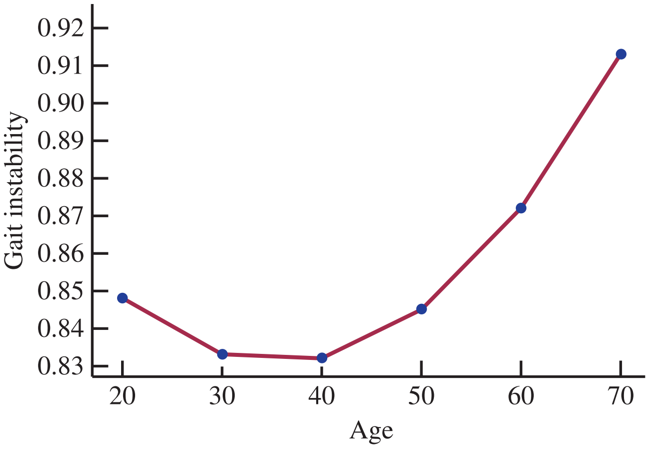
14.1 **a.** A deterministic model does not have the random deviation component e, while a probabilistic model does contain such a component.

14.3 **a.** (mean y value for fixed values of $x1$, $x2$, $x3$) $= 30 + 0.90x_1 + 0.08x_2 - 4.5x_3$
b. $\beta_0 = 30$, $\beta_1 = 0.9$, $\beta_2 = 0.08$, $\beta_3 = -4.50$ **c.** The average change in acceptable load associated with a 1-cm increase in left lateral bending, when grip endurance and trunk extension ratio are held fixed, is 0.90 kg. **d.** The average change in weight associated with a 1 N/kg increase in trunk extension ratio, when grip endurance and left lateral bending are held fixed, is −4.5 kg.

14.5 **a.** 13.552 g. **b.** When length is fixed, the mean increase in weight associated with a 1-mm increase in width is 0.828 g. When width is fixed, the mean increase in weight associated with a 1-mm increase in length is 0.373 g.

14.7 **a.** 35,893 **b.** $\beta_1 = 0.974$ is the change in power consumption associated with an increase of 1 million in population when gross domestic product is held constant.

14.9 **a.**



b. From 20 to 30, decreases by 0.15. From 50 to 60 increases by 0.027.

14.11 **c.** The parallel lines in each graph are attributable to the lack of interaction between the two independent variables.

14.13 **a.** $y = \alpha + \beta_1x_1 + \beta_2x_2 + \beta_3x_3 + e$
b. $y = \alpha + \beta_1x_1 + \beta_2x_2 + \beta_3x_3 + \beta_4x_1^2 + \beta_5x_2^2 + \beta_6x_3^2 + e$
c. $y = \alpha + \beta_1x_1 + \beta_2x_2 + \beta_3x_3 + \beta_4x_1x_2 + e$;
$y = \alpha + \beta_1x_1 + \beta_2x_2 + \beta_3x_3 + \beta_4x_1x_3 + e$;
$y = \alpha + \beta_1x_1 + \beta_2x_2 + \beta_3x_3 + \beta_4x_2x_3 + e$
d. $y = \alpha + \beta_1x_1 + \beta_2x_2 + \beta_3x_3 + \beta_4x_1^2 + \beta_5x_2^2 + \beta_6x_3^2 + \beta_7x_1x_2 + \beta_8x_1x_3 + \beta_9x_2x_3 + e$

Copyright 2025 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s). Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

14.15 **a.** Three dummy variables would be needed to incorporate a nonnumerical variable with four categories. For example, you could define $x_3 = 1$ if the car is a subcompact and 0 otherwise, $x_4 = 1$ if the car is a compact and 0 otherwise, and $x_5 = 1$ if the car is a midsize and 0 otherwise. The model equation is then $y = \alpha + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + \beta_5 x_5 + e$. **b.** For the variables defined in Part (a), $x_6 = x_1 x_3$, $x_7 = x_1 x_4$, and $x_8 = x_1 x_5$ are the additional predictors needed to incorporate interaction between age and type of car.

14.17 **a.** $0.01 < P\text{-value} < 0.05$ **b.** $P\text{-value} > 0.10$ **c.** $P\text{-value} = 0.01$ **d.** $0.001 < P\text{-value} < 0.01$

14.19 **a.** $F = 12118$, P-value ≈ 0, reject H_0 **b.** Since the P-value is small and r^2 is close to 1, there is strong evidence that the model is useful. **c.** The model in Part (b) should be recommended, since adding the variables x_1 and x_2 to the model [to obtain the model in Part (a)] only increases the value of R^2 a small amount (from 0.994 to 0.996).

14.21 $F = 24.41$, P-value < 0.001, reject H_0 and conclude that the model is useful.

14.23 **a.** Note that x_3 is an indicator variable. Therefore, the value of the coefficient can be interpreted as follows: The estimated value of $b_3 = -84.6$ is a "correction" or "adjustment" that we would make based on the type of printer to the mean Price of the printer, with the other variables remaining constant. The Price for a laser printer is, on average, \$84.6 less than that for an inkjet, with the other variables (ink cost per year and text print speed) remaining constant. **b.** 42.40% (this is the adjusted R^2 value)

14.25 **a.** $\hat{y} = 1.44 - 0.0523\,length + 0.00397\,speed$ **b.** 1.3245 **c.** $F = 24.02$, P-value ≈ 0, reject H_0 and conclude that the model is useful. **d.** $\hat{y} = 1.59 - 1.40\left(\frac{\text{length}}{\text{speed}}\right)$
e. The model in Part (a) has $R^2 = 0.75$ and R^2 adjusted $= 0.719$, whereas the model in Part (d) has $R^2 = 0.543$ and R^2 adjusted $= 0.516$.

14.27 $F = 8.56$, P-value $= 0.0004$, reject H_0. There is evidence that the quadratic model is useful.

Chapter 15

15.1 **a.** 0.007 **b.** 0.163 **c.** 0.010 **d.** 0.000 **e.** 0.074 **f.** 0.019

15.3 **a.** H_0: $\mu_1 = \mu_2 = \mu_3 = \mu_4$, H_a: At least two of the four μ_i's are different. **b.** P-value $= 0.012$, fail to reject H_0 **c.** P-value $= 0.012$, fail to reject H_0

15.5 $F = 3.124$, P-value $= 0.038$, reject H_0. There is sufficient evidence to conclude that the mean rating associated with the game description by 7- to 8-year-old children is not the same for all four restrictive rating labels.

15.7 **a.** $F = 3.86$, P-value $= 0.035$, reject H_0.There is sufficient evidence to conclude that the mean hunger rating is not the same for all three treatments ("healthy" snack, "tasty" snack, no snack). **b.** It is not reasonable to conclude that the mean hunger rating is greater for people who do not get a snack because the sample mean hunger rating is highest for the healthy snack group.

15.9 $F = 21.356$, P-value ≈ 0, reject H_0. There is sufficient evidence to conclude that the mean mass for this species of lizards is not the same for all four island lizard populations.

15.11

a. H_0: $\mu_1 = \mu_2 = \mu_3 = \mu_4$
H_a: At least two among: μ_1, μ_2, μ_3, μ_4 are different
b. Treatment df $= k - 1 = 4 - 1 = 3$
Error df $= N - k = 585 - 4 = 581$
c. Given that the researchers drew the conclusion that there was no significant improvement in episodic memory, a P-value larger than the significance level is required so that the null hypothesis is not rejected. The F statistic of 1.5, with treatment and error degrees of freedom of 3 and 581, respectively, resulted in a P-value of 0.27, and the other two F statistics yielded P-values of essentially zero. Therefore, the F statistic could be 1.5, and not 11.5 or 111.5.

15.13

```
Source of        df       Sum of          Mean      F
Variation                Squares        Square

Treatments        3    75,081.72     25,027.24   1.70
Error            16   235,419.04     14,713.69
Total            19   310,500.76
```

$F = 1.70$, P-value > 0.10, fail to reject H_0

15.15 Since there is a significant difference in all three of the pairs we need a set of intervals that does not include zero. Set 3 is therefore the required set.

15.17 **a.** There were no significant differences between the sugar content for raw and boiled potatoes, as well as for boiled and microwaved potatoes. However, there was a significant difference between the mean sugar content of potatoes that were raw and those that were microwaved or fried, and a significant difference between potatoes that were fried and those that were boiled or microwaved. **b.** Answers will vary. One noticeable difference is that, relative to the other preparation methods, fried potatoes move from one of the lower starch content methods to the highest sugar content method. Starch content also had more significant differences between preparation methods than the sugar content.

15.19 **a.**
$\mu_1 - \mu_2$: $(-3.87, -1.91)$
$\mu_1 - \mu_3$: $(-2.48, -0.58)$
$\mu_1 - \mu_4$: $(-1.86, 0.04)$
$\mu_2 - \mu_3$: $(0.38, 2.34)$
$\mu_2 - \mu_4$: $(1.00, 2.96)$
$\mu_3 - \mu_4$: $(-0.33, 1.57)$

b. The underscoring pattern is shown below. The T-K confidence intervals computed in Part (a) show that Bimini and New Providence Islands have an interval that contains zero, as does New Providence and Harbor Islands. The underscoring pattern indicates no significant difference in lizard weights between Bimini and New Providence Islands, and no significant difference in lizard weights between New Providence and Harbor Islands. The other island pairs have significant differences.

	Bimini	New Providence	Harbor	Great Exuma
Sample Mean	4.22	5.13	5.75	7.11

15.21 The underscoring pattern indicates that there is no significant difference in mean inflammatory response for diacetyl between Control and Low Dose, and High Dose is significantly different from both Control and Low Dose.

Index

1 in *k* systematic sample, 42

A

A *and* B, 308
A *or* B, 308
Addition rule
  for mutually exclusive events, 321–322
  general, 345–346
Additive multiple regression model, 770
Additive probabilistic model, 729
Adjusted R^2, 786
Alternative hypothesis, 546–547
Analysis of variance (ANOVA). *See* Single-factor analysis of variance (ANOVA)
ANOVA table, 805

B

Bar chart
  for categorical data, 13
  comparative, 80–81
  uses of, 86–88
Bayes' Rule, 351–353
Bell-shaped curve, 113. *See also* Normal distribution
Bias in sampling, 34–37
Biased statistic, 495
Bimodal histogram, 112
Binomial distribution, 403–409
  defined, 403
  formula for, 404–405
  mean of, 409
  normal approximation to, 448–450
  standard deviation of, 409
Binomial experiment, 403
Binomial random variable, 403
Bivariate data
  analysis, 289–290
  cautions and limitations, 290–292
  scatterplot of, 122
Bivariate data set, 10
Blocking, 50
  defined, 47
  extraneous variables and, 47
  random assignment and, 57
Bootstrap confidence interval
  defined, 528, 535
  for difference between two population means
    using independent samples, 663–665
    using paired samples, 665–667
  for difference between two population proportions or two treatment proportions, 670–671
  for difference between two treatment means, 662
  for one proportion, 529–532
  for a population mean, 534–536
Bootstrap distribution, 528, 535
Boxplots, 194–196
  comparative, 195
  modified, 196
  outlier and, 195
  skeletal, 194, 195

C

Categorical data, 9, 180–181
  bar chart for, 13
  chi-square test for, 692–696
  comparative bar chart for, 80–81
  frequency distribution for, 11–13
  multivariate, 712
  pie chart for, 81–84
Causation, correlation and, 231
Cell count. *See* Expected cell count; Observed cell count
Census, 34
Center of data set, 174–181
Central Limit Theorem, 473–475
Chance experiment, 306–307
Chebyshev's rule, 201–203
Chi-square distribution, 695
Chi-square tests
  goodness-of-fit, 696–699
  homogeneity, 702–712
  independence, 708–709
  for univariate data, 692–699
Class interval, 106, 108
Cluster sampling, 41–42
Clusters, 41
Coefficient of determination, r^2, 256–259
Coefficient of multiple determination, R^2, 785
Common population proportion, 646

Comparative bar chart, 80–81
Comparative boxplots, 195
Comparative stem-and-leaf display, 97–98
Complement of an event, 308
Complete second order model, 775
Completely randomized design, 57
Conditional probability, 325–333
  compared to unconditional probability, 329, 349
Confidence interval
  for comparing two population or treatment means using
    independent samples, 625
    paired samples, 638–640
  for comparing two population or treatment
      proportions, 650–652
  general form of, 505–506
  for population mean, 515
  for population proportion, 501–505
  for the slope of the population regression line, 743
Confidence level, 499
Confounding variables, 30, 47
Contingency table, 702
Continuity correction, 448
Continuous numerical data, 10
  defined, 11
  frequency distribution, 106–108
  histogram, 108–110
Continuous random variable, 375, 385
  mean of, 395
  probability distribution of, 385–388
  standard deviation of, 395
Control group, 54, 62–63
Convenience sampling, 42
Correlation, 224–231
  causation and, 231
Correlation coefficient, 224–225
  cautions and limitations, 290–292
  coefficient of determination and, 258
  formula for, 225
Cumulative area, 387
Cumulative relative frequency, 116–117
Cumulative relative frequency plot, 117–118
Curvature in residual plot, 252

D

Danger of extrapolation, 239–240, 735
Data, types of, 9–10
Data analysis process, 6–7
Degrees of freedom (df), 188
  ANOVA and, 801
  chi-square distributions and, 695
  simple linear regression and, 736
  t distribution and, 513–514
  two-sample *t* test and, 617
Density, 110
Density curve, 385
Density function, 385
Density scale, 110
Dependent events, 338
Dependent variable, 235
Descriptive statistics, 7
Deterministic relationship, 729
Deviations from the mean, 184–186
Diagram of experimental design, 54
Direct control, 47, 50
Discrete data, 10
  defined, 11
  frequency distribution, 102–106
  histogram, 102–106
Discrete probability distribution
  binomial distribution, 403–409
  geometric distribution, 410–411
  properties of, 380
Discrete random variable
  defined, 375
  mean value of, 391–393
  probability distribution of, 378–383
  standard deviation of, 393–395
  variance of, 394
Disjoint, 309
Disjoint events. *See* Mutually exclusive events
Dotplot, 14–16
Double-blind experiment, 63–64

E

Empirical Rule, 203–204
  standard normal distribution and, 422, 425
Equally likely outcomes, 313
Error sum of squares (SSE), 801
Errors in hypothesis testing, 550–553
Estimated standard deviation, 736
Estimation
  bias in, 495
  choosing a statistic, 494–497
  interval estimate (*See* Confidence interval)
  point estimate, 493
Event, 307–311
  foming new events, 308–310
  simple, 307
Exact binomial test
  for population proportion, 594–597
Expected cell count, 705
Expected value. *See* Mean
Experiment, 29, 30, 46
  completely randomized, 57
  double-blind, 61–62
  replicating, 49–50
  single-blind, 63–64
  using a control group, 54, 62
  using volunteers, 42, 65
Experimental conditions, 46
Experimental design
  evaluation of, 52–54
  principles of, 50
Experimental unit, 64–65

Copyright 2025 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s). Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

Explanatory variables, 46, 50, 235
Extraneous variables, 47–48, 51
Extrapolation, danger of, 239, 240, 735
Extreme outlier, 195

F

F distribution, 786–787
F test
  ANOVA, 798–806
  for model utility, 787–792
Factor, 29, 30, 46, 798
Finite population correction factor, 525–526
Fitted value. *See* Predicted value
Five-number summary, 194
Frequency, 12
Frequency distribution, 11
  for categorical data, 11–13
  for continuous numerical data, 106–108
  for discrete numerical data, 102–106
Full quadratic model, 775
Fundamental identity for single-factor ANOVA, 804

G

General addition rule, 345–346
General additive multiple regression model, 770
General multiplication rule, 347–349
Geometric probability distribution, 410
Geometric random variable, 410
Goodness-of-fit test, 696–699
Goodness-of-fit statistic, 695
Grand total, 702
Graphical data display
  cautions and limitations, 153–156

H

Heavy-tailed distribution, 113
Histograms
  continuous numerical data, 108–110
    with equal width intervals, 108
    with unequal width intervals, 110
  discrete numerical data, 102–106
  probability, 380, 385, 391, 408–409
  shapes, 112–113
  using density, 110–111
Homogeneity, chi-square test of, 702–712
Hypergeometric distribution, 407
Hypotheses, 546–548
  alternative hypothesis, 546
  null hypothesis, 546
Hypothesis test
  comparing two population or treatment means using
    independent samples, 614–627
    paired samples, 632–640
  errors in, 550–553
  for comparing two population or treatment proportions, 645–652
  for goodness-of-fit, 696
  for homogeneity, 705
  for independence, 705
  for a population mean, 568–575
  for a population proportion, 556–566
  for the slope of the population regression line, 741–749
  power of (*See* Power)
  randomization test
    for one mean, 599–600
    for one proportion, 590–594
    for two means, 660–667
    for two proportions, 670–673
  significance level of, 552
  steps for, 564

I

Implicit null hypothesis, 547
Independence, 337–342
  testing for, 709
Independent events
  defined, 338
  multiplication rule for more than two, 340
  multiplication rule for two, 338–339
Independent samples, 614–616, 632
  comparing two population or treatment means using, 614–632
  simulation-based inference for difference in two population means using, 663–665
Independent variable, 235
Indicator variable, 777
Inferential statistics, 7
Influential observation, 255, 756
Interaction, 773–777
  predictor in multiple regression, 778–779
Intercept, 236
Interquartile range, 188–189
Intersection of two events, 308
Interval estimate. *See* Confidence interval

J

Joint probability table, 337

L

Large sample confidence interval
  for the difference in two population or treatment proportions, 650–651
  for a population proportion, 498–507
Large-sample hypothesis test
  for the difference in two population or treatment proportions, 646–652
  for a population proportion, 556–566

Copyright 2025 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s). Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

Law of large numbers, 316
Law of total probability, 349–351
Leaf, 94
Least squares regression line
  assessing fit of, 248–264
  defined, 237
  deviations from, 236–238
  influential observations and, 255
  making predictions with, 239
  outliers and, 255
  population regression line and, 733
  slope of, 237, 238
  standard deviation about, 259–264
  y intercept of, 237
Least-squares estimates, 783
Left skewed histogram, 113
Light-tailed distribution, 113
Line
  assessing fit of, 248–264
  equation of, 236
Linear combination of random variables, 397, 399–400
  mean and standard deviation of, 396–399
Linear function of random variable, 396, 399–400
  mean and standard deviation of, 395
Linear regression, 262
  model, 730–733
  model utility test for simple linear regression, 746
Linear relationship, strength of, 224–225, 228
Lower quartile, 188–189
Lower tail of histogram, 113
Lower-tailed test, 562, 571

M

Margin of error
  defined, 506
  sample size choice and, 506–507, 518
Marginal total, 702
Mean
  of binomial distribution, 409
  of a continuous random variable, 395
  of a data set, 174–178
  defined, 174
  deviations from, 184–186
  of discrete random variable, 391–393
  of a linear combination, 397
  of a linear function, 395–396
  of normal distribution, 414
  outliers and, 177
  population mean, 176
  of random variable, 389
  sample mean, 175
  of sampling distribution, 469–477
  vs. median, 179–180
Mean square, 801
  for error (MSE), 801–803
  for treatments (MSTr), 801–803
Measure of center, 174–181
Measure of relative standing, 204–205
Measurement bias, 35, 36
Median
  of a data set, 178–179
  defined, 178
  outliers and, 179
  sample median, 178
  vs. mean, 179–180
Mild outlier, 195
Mode, 112
Model equation, 729
Model utility test
  in multiple regression, 787–788
  for simple linear regression, 746
Modified boxplots, 196
Multimodal histogram, 112
Multiple comparison procedure, 810–814
Multiple regression
  general additive model, 770
  interaction variables in, 773–777
  model utility *F* test, 787–788
  qualitative predictors in, 777–778
Multiplication rule
  general, 347–349
  for independent events, 338–340
Multivariable graphical displays, 131–141
Multivariate data, 10
Mutually exclusive events, 309–310
  addition rule for, 321

N

Negative skew, 113
Nonlinear regression, 270–285
  choosing models, 282–285
  common nonlinear functions, 270–272
Nonresponse bias, 35–36
Normal approximation to the binomial distribution, 448–449
Normal curve, 113. *See also* Normal distribution
  discrete variables and, 447–448
Normal distribution, 414–433
  extreme values in, 432–433
  mean value of, 414
  standard deviation of, 414
  standard normal distribution, 414–418
Normal probability plot, 437–438
Normal score, 437
Null hypothesis, 546, 547–548
Numerical data, 9
  continuous, 10–11
  discrete, 10–11
  displaying bivariate, 122–128
  dotplots for, 14–16

Copyright 2025 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s). Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

Numerical data (*Continued*)
  frequency distribution for
    continuous numerical data, 106–107
    discrete numerical data, 102–106
  histograms for
    continuous numerical data, 108–110
    discrete numerical data, 102–106
  stem-and-leaf display for, 94–98
Numerical descriptive measures
  cautions and limitations, 211–212

O

Observational study, 29–31
Observed cell count, 702, 704
Observed significance level, 558. *See also* *P*-value
One sample z confidence interval, 511
One-sample t confidence interval, 514–519
One-sample t test, 571
One-way frequency table, 692
Outlier, 95, 177, 179, 195, 1255, 756
Overcoverage, 67

P

Paired data, 16
Paired samples, 615, 633
  comparing two population or treatment means using, 632–640
  paired t test and, 634–638
  simulation-based inference about two population means using, 665–667
Paired t confidence interval, 638–640
Paired t test, 634–638
Pearson's sample correlation coefficient. *See* Correlation coefficient
Percentile, 205–206
Pie chart
  for categorical data, 81–84
  uses of, 86–88
Placebo, 63
Plotting residuals, 252–256, 756–761
Point estimation, 493–497
Polynomial regression, 771–773
Pooled t test, 624
Population, 6, 7
Population correlation coefficient, 231
Population mean, 176
  bootstrap confidence intervals for, 534–536
  comparing two using independent samples, 614–627
  comparing two using paired samples, 632–640
  confidence interval for, 515
  hypothesis test for, 568–575
  one-sample t test for, 571
  randomization test for, 590–592
  simulation-based inference for two means, 660–667
Population proportion
  bootstrap confidence interval for, 527–532
  comparing two using large samples, 646–650
  exact binomial test for, 594–597
  large-sample confidence interval for, 498–507
  large-sample hypothesis test for, 556–566
  randomization test for, 590–594
  simulation-based inference for two proportions, 670–673
Population proportion of S's, 181
Population regression coefficients, 770
Population regression function, 771, 773
Population regression line, 730–733
Population standard deviation, 188
Population variance, 188
Positive skew, 113
Potentially influential observation, 756
Power
  calculating, 581–582
  defined, 579
  effect of sample size and significance level on, 579–580
  type II error probabilities and, 578–585
Practical significance, 574, 589
Predicted value, 248–251, 784
Predictor variable, 235
Principle of least-squares, 236–242
Probabilistic model, 729
Probability
  of an event, 313, 316, 318–319
  basic properties of, 318–322
  classical approach for equally likely outcomes, 313–314
  of the complement of an event, 340
  conditional, 325–333
  estimating empirically, 357–360
  estimating using simulation, 360–365
  histogram, 380, 385, 391, 408–409
  of the intersection of two events, 338, 347
  relative frequency approach to, 314–317
  subjective approach to, 317–318
  of the union of two events, 321, 345
Probability distribution
  for continuous random variable, 385–388
  for discrete random variable, 378–383
Probability histogram, 380, 385, 391, 408–409
P-value
  alternative hypotheses and, 563
  defined, 558
  determining for a chi-square test, 696
  determining for a t test, 571
  determining for a z test, 560–566

Q

Quadratic regression, 273–277, 771, 772
Qualitative data, 9
Qualitative predictor variables, 777–778
Quantitative data, 9
Quartile
  lower, 188–189
  upper, 188–189

Copyright 2025 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s). Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

R

r. See Correlation coefficient
Random assignment, 47–52
Random deviation, 730, 738
Random number, 38
Random sampling, 37
Random selection, 39
Random variable, 375–377
  linear combination of, 397
  linear function of, 396
  mean value of, 389
  standard deviation of, 389
Randomization distribution, 590
Randomization test
  for a difference between two means, 660–667
  for a difference between two proportions, 671–673
  for a population mean, 599–600
  for a population proportion, 590–594
Randomized block design, 58
Range, 184
Re-expression. *See* Transformations
Regression. *See* Least squares regression line; Linear regression
Regression analysis, 242–243
  linear regression, 262
  nonlinear regression, 270–285
Regression sum of squares (SSRegr), 787–788
Relative frequency, 12
Relative frequency distribution, 12
Replacement, sampling with and without, 38–39, 340–342, 407–408
Replication, 49, 50, 64–65
Resampling, 528
Residual
  in linear regression, 248–251
  in multiple regression, 784
  plotting, 252–256, 756–761
Residual analysis, 753–755
Residual plot, 252–256, 756–761
  defined, 756
  standardized residual plot, 756–758
Residual sum of squares (SSResid), 256–258, 785
Response bias, 35, 36
Response variable, 46, 50, 235
Right skewed histogram, 113
Robust methods, 761
*r*th percentile, 205

S

Sample, 6, 7
  independent samples, 615, 632
  paired samples, 632–640
Sample correlation coefficient, 225–226
  properties, 228–229
Sample mean, 175
Sample median, 178
Sample proportion of successes, 180–181
Sample regression line, 237, 242. *See also* Least squares regression line
Sample size, 39–40
  margin of error and, 506–507, 518
  power and, 579–580
Sample space, 306–307
Sample standard deviation, 186
Sample variance, 186
Sampling, 34–45
  bias in, 34–37
  cluster, 41–42
  convenience, 42
  random, 37
  stratified, 40–41
  systematic, 42
  with and without replacement, 38–39, 340–342, 407–408
Sampling distribution, 464, 467
  of the difference in sample means (independent samples), 615
  of the difference in sample proportions, 646
  of the sample mean, 469–477
  of the sample proportion, 479–483
  of the slope of the least squares regression line, 742
Sampling frame, 38
Sampling variability, 114, 465–467
Scatterplot, 122
Segmented bar charts, 84–86
Selection bias, 35, 36
Significance level, 552, 560
  power and, 579–580
Simple comparative experiment, 46–62
  design strategies for, 47
Simple event, 307
Simple linear regression. *See* Least squares regression line; Linear regression
Simple random sample, 37
  selection of, 38–39
Simulation to estimate probabilities, 360–365
Simulation-based inference
  for two means, 660–667
  for two proportions, 670–673
Simultaneous confidence level, 812
Single-blind experiment, 64
Single-factor analysis of variance (ANOVA), 798
  F test for, 798–806
  fundamental identity for, 804
Skeletal boxplots, 194
Skewed histogram, 113
Slope
  defined, 236
  estimate for population regression line, 741–749
  of least-squares regression line, 237, 238, 291
Smoothed histogram, 112
Stacked bar chart. *See* Segmented bar charts

Standard deviation
  about the least-squares line, 259–264
  of the binomial distribution, 409
  of a continuous random variable, 395
  of a data set, 186–188
  of a difference in means, 615
  of a discrete random variable, 393–395
  of a linear combination, 397
  of a linear function, 395–396
  of a normal distribution, 414
  of a random variable, 389
  of the slope of the least-squares line, 741
Standard error, 506, 523
Standard normal distribution, 415–418
Standardized residual, 753
Standardized residual plot, 756–758
Standardizing, endpoints, 428
Statistical significance, 574–575
Statistics
  defined, 1, 465
  descriptive, 7
  inferential, 7
Stem-and-leaf display, 94–98
  comparative, 97–98
  repeated stems, 96–97
Stems, 94
Strata, 40
Stratified random sampling, 40–41
Studentized range distribution, 810
Sum of the squared deviations, 237
Sum of squares
  error (SSE), 801
  regression (SSReg), 787–788
  total (SSTo), 256–258, 785, 804
  treatment (SSTr), 801
Symmetric histogram, 112
Systematic sampling, 42

T

t confidence interval
  for a difference in population or treatment means, 625
  for a population mean, 511
t critical value, 514
t distribution, 513–514
  degrees of freedom and, 513
  properties of, 513
t test
  finding *P*-values for, 570–571
  one-sample *t* test for a population mean, 568–575
  paired *t* test, 634–638
  pooled *t* test, 624
  power of and type II error probabilities, 581–583
  for the slope of the population regression line, 745
  two-sample *t* test for difference in population or treatment means, 614–627
Test of hypotheses. *See* Hypothesis test
Test procedure, 546, 550
Test statistic, 557, 558
Time series plot, 125–126
Total sum of squares (SSTo), 256–258, 785, 804
Transformations, 275–277
  defined, 277
  models transforming *x,* 277–280
  models transforming *y,* 280–282
  nonlinear regression models, 270–285
Transforming data to normal distribution, 438–442
Treatment, 46, 50
Treatment means
  comparing using independent samples, 621–624
  comparing using paired samples, 632–640
  confidence interval for comparing using independent samples, 625
  confidence interval for comparing using paired samples, 638–640
  simulation-based inference about the difference in, 660–662
Treatment proportions
  bootstrap confidence intervals for the difference in, 670–671
  confidence interval for difference in, 650–651
  large sample test for difference in, 646–650
  randomization tests for the difference in, 671–673
Treatment sum of squares (SSTr), 801
Tree diagram, 307
Tukey-Kramer (T-K) multiple comparison procedure, 810–813
  simultaneous confidence level and, 812
Two-sample *t* confidence interval, 625
Two-sample *t* test
  for comparing two population means, 617
  for comparing two treatment means, 622
Two-tailed test, 562, 571
Two-way frequency table, 702–703
Type I error, 550–553
  probability of, 552
Type II error, 550–553
  power and, 578–585
  probability of, 552, 578–585

U

Unbiased statistic, 495–496
Unconditional probability
  compared to conditional probability, 329, 349
Undercoverage, 35, 67
Unexplained variability, 257
Uniform distribution, 386
Unimodal histogram, 112–113
Union of two events, 321
Univariate data set, 9
Upper quartile, 188–189
Upper tail of a histogram, 113
Upper-tailed test, 562, 570

Copyright 2025 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s). Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

V

Variability
  in data set, 184–191
  interpreting, 200–206
  role of, 3–5
  sampling, 465–467
  unexplained, 257
Variables, 9
  blocking, 47
  confounding, 30, 47
  continuous random. (*See* Continuous random variable)
  dependent, 235
  discrete random. (*See* Discrete random variable)
  explanatory, 46, 50, 235
  extraneous, 47–48, 51
  independent, 235
  indicator, 777
  interaction between, 773–777
  predictor, 235
  random, 375–377
  response, 46, 50, 235
Variance, 186–188. *See also* Standard deviation
  of discrete random variable, 394
Venn diagram, 309–310
Vertical intercept, 236
Voluntary response sampling, 42
Volunteers in experiments, 42, 65

W

Whiskers in a boxplot, 194–196

Y

y-intercept, 236
  of the least-squares regression line, 237, 291
  of the population regression line, 731

Z

z confidence interval
  for a difference in population or treatment proportions, 650–651
  for a population mean, 511
  for a population proportion, 505
z critical value, 502, 503, 504
z curve, 415. *See also* Standard normal distribution
z test
  for a difference in population proportions, 647
  for a population proportion, 563
z-score, 204, 428

Copyright 2025 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s). Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.